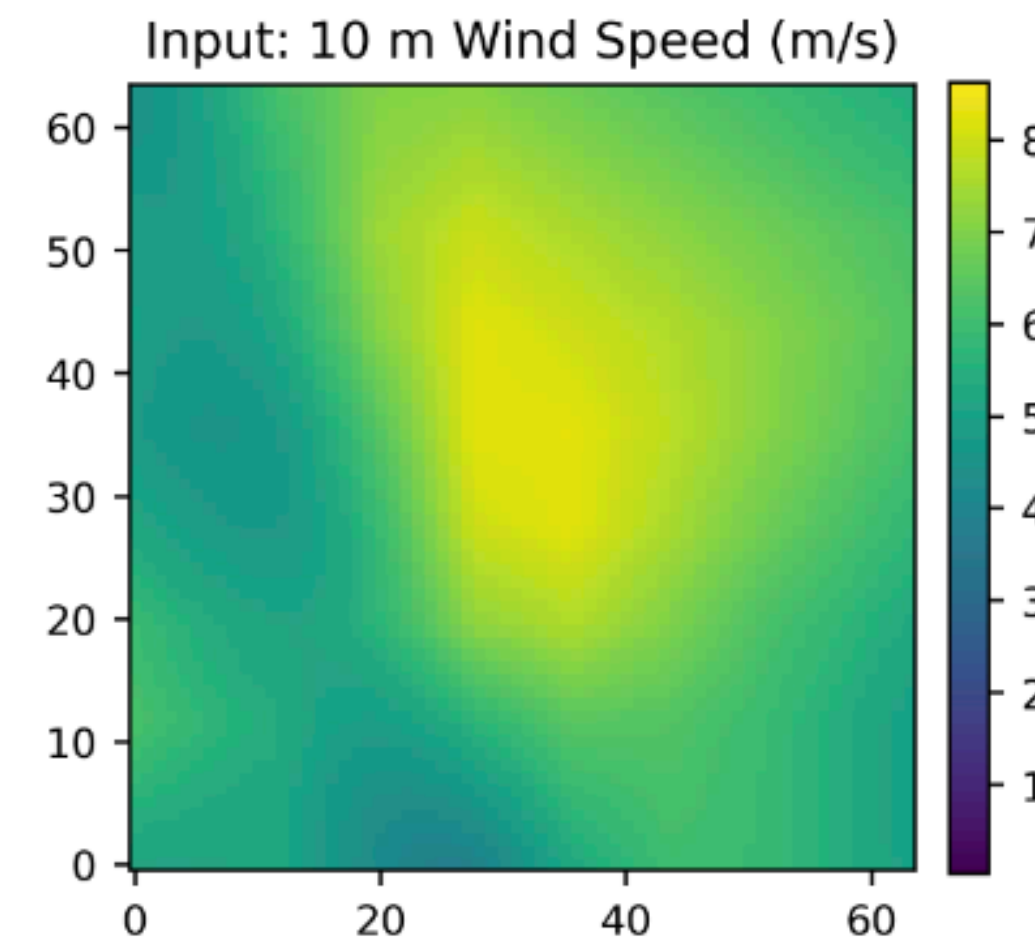




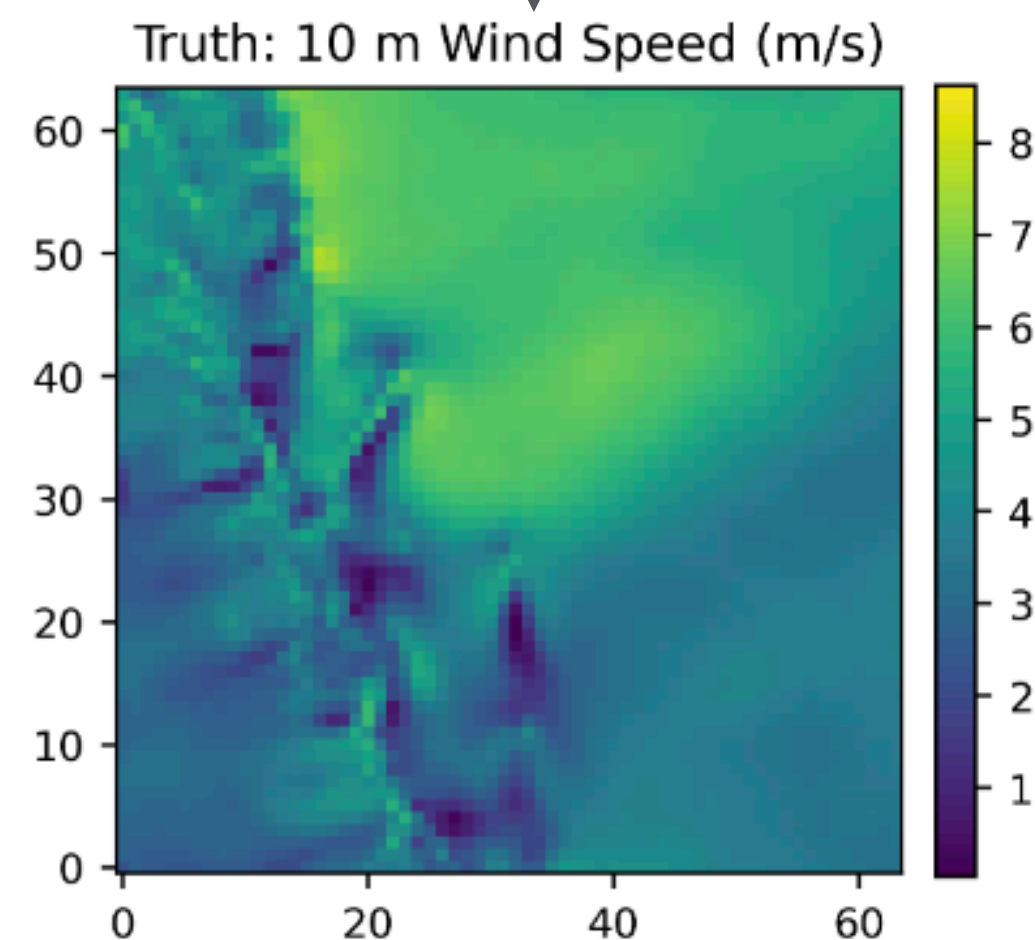
TOWARDS LARGE SCALE CIRCUIT SIMULATION FOR HYBRID DIFFUSION MODEL

WHAT STATISTICAL DOWNSCALING MUST RECOVER

- Statistical downscaling is a computational technique used to enhance the spatial resolution of coarse geospatial data, making it usable for local scale applications
- But it is not only super-resolution; it is conditional reconstruction of fine-scale structure from a large-scale state
- This motivates models that can represent interactions across spectral groups, channels, or scales

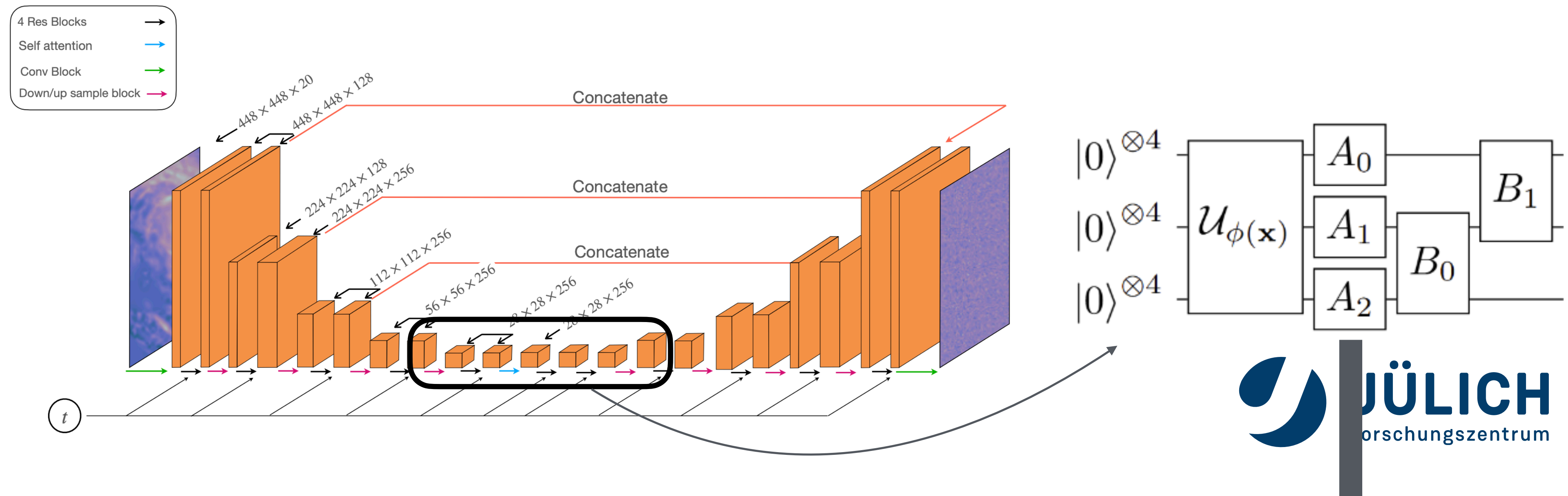


Cross-group interaction:
sharp gradients
local extremes
anisotropic fine structure



RECAP FOR QR4EO PROJECT

- We tested a hybrid quantum-classical diffusion model based on the state-of-the-art CorrDiff (@Nvidia) residual collective diffusion model
- The CorrDiff framework is based on a two-step residual corrective diffusion approach for statistical downscaling, which composed of a regression UNet and a diffusion UNet
- We quantumized the diffusion UNet by replacing parts of the latent channels in the bottleneck to quantum circuits



RECAP FOR QR4EO PROJECT

- Four qubits encode one 2x2 (spatial resolution) latent channel
- We have tested several quantum circuits:
 - A block only, contains only inter-channel interaction
 - B block only, contains only sparse inter-channel interaction
 - A+B block

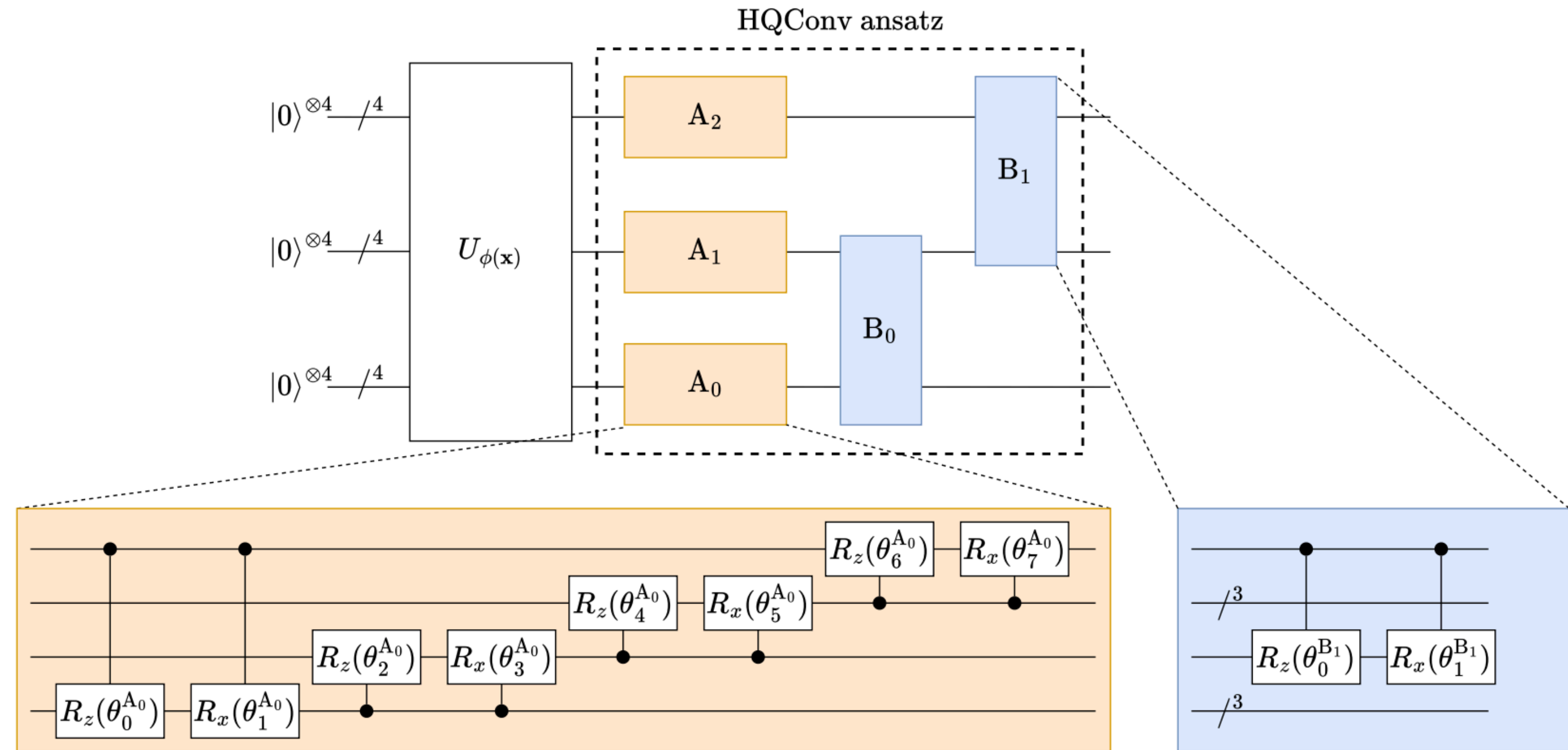
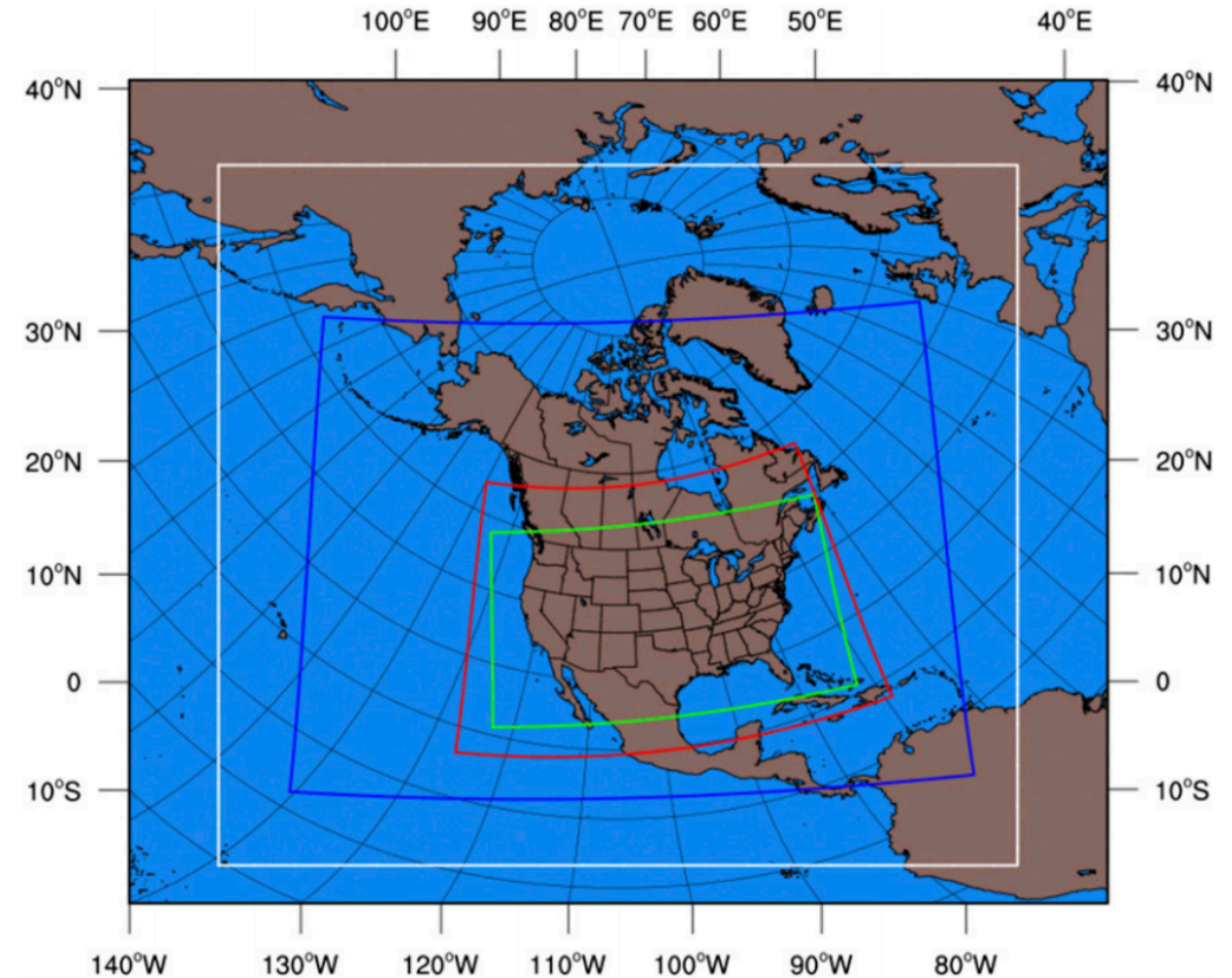


Figure 5: HQConv ansatz proposed in [29] initially extracts local information, then uses additional controlled gates to encode intra-channel information.

THE DATASET: HRRR-MINI

- A reduced dataset derived from ERA5 (low-resolution) reanalysis and the corresponding HRRR (high-resolution) analysis fields, covering the continental United States and surrounding regions
- Low-resolution meteorological fields are used as conditioning inputs, and the model reconstructs high-resolution 10 m wind fields
- The data are collected from the period 2018-2021
- The target here: 10 m zonal (u_{10m}) and meridional (v_{10m}) wind components



Training setup

Training: 2018-2020
Testing: 2021

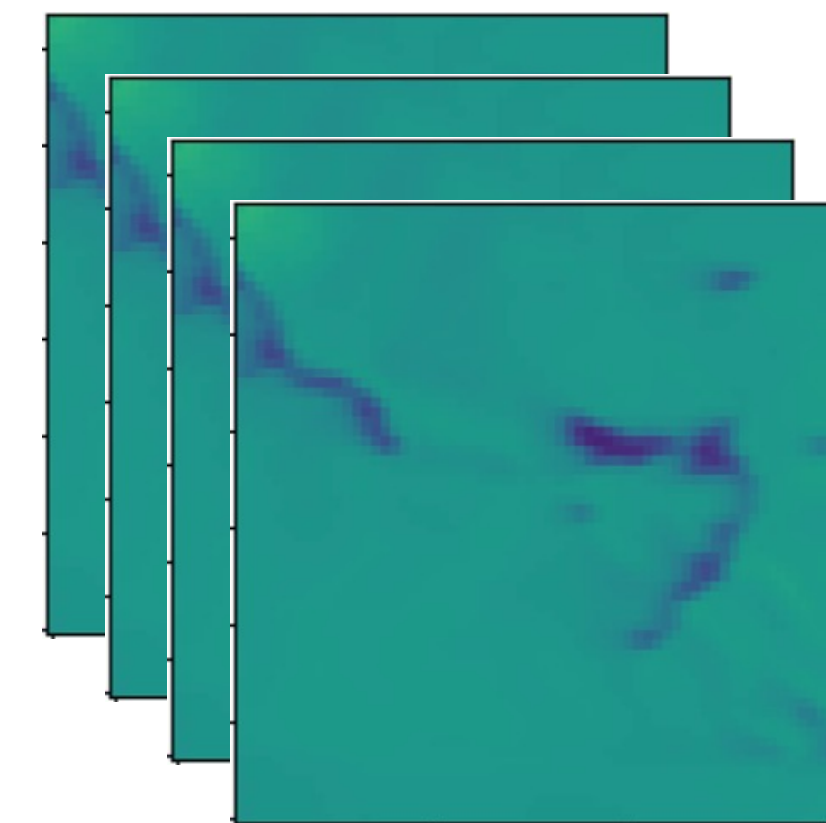
HRRR-mini (NVIDIA): 100K low-high res pairs from continental US (2018-2021)



- Temperature (t)
- Geopotential height (z)
- U and V wind components
- Specific humidity (q)
- Surface-level: 10m winds, 2m temperature, total column water vapor, surface pressure, mean sea level pressure

**ERA5 reanalysis
(ECMWF)**

31 km resolution



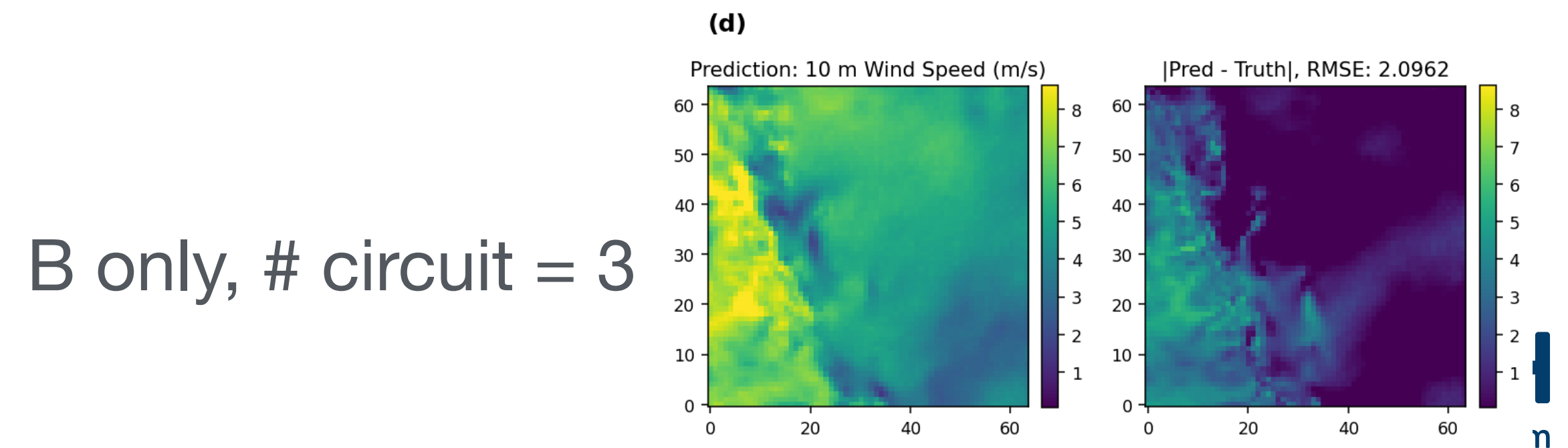
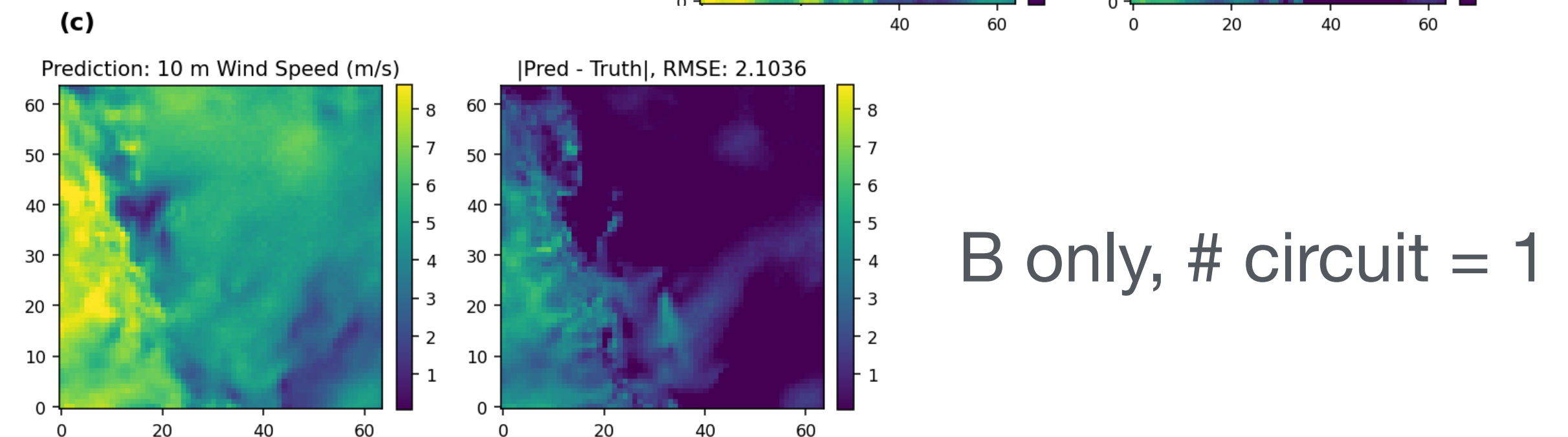
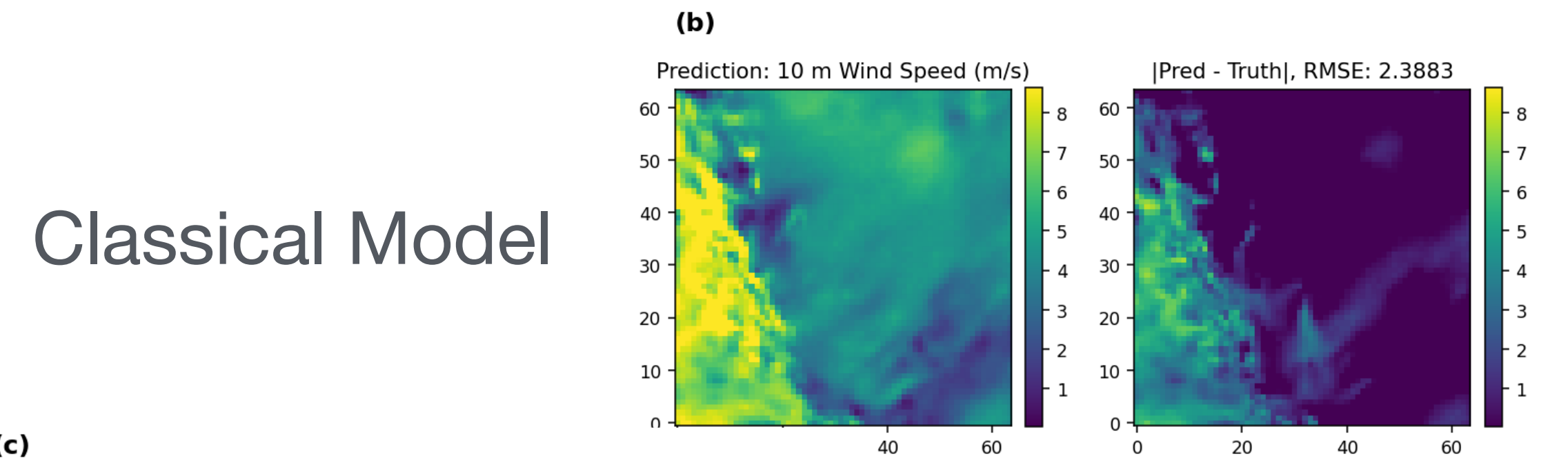
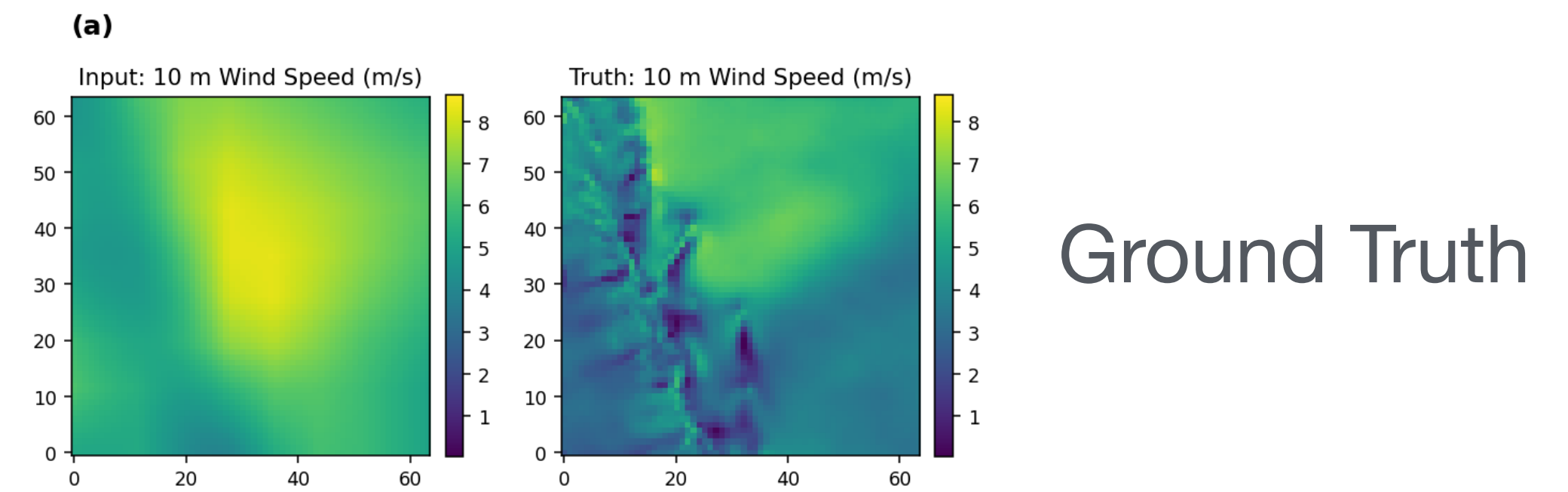
- 10m winds
- 2m temperature
- Total precipitation

**HRRR analysis
(NOAA)**

3 km resolution

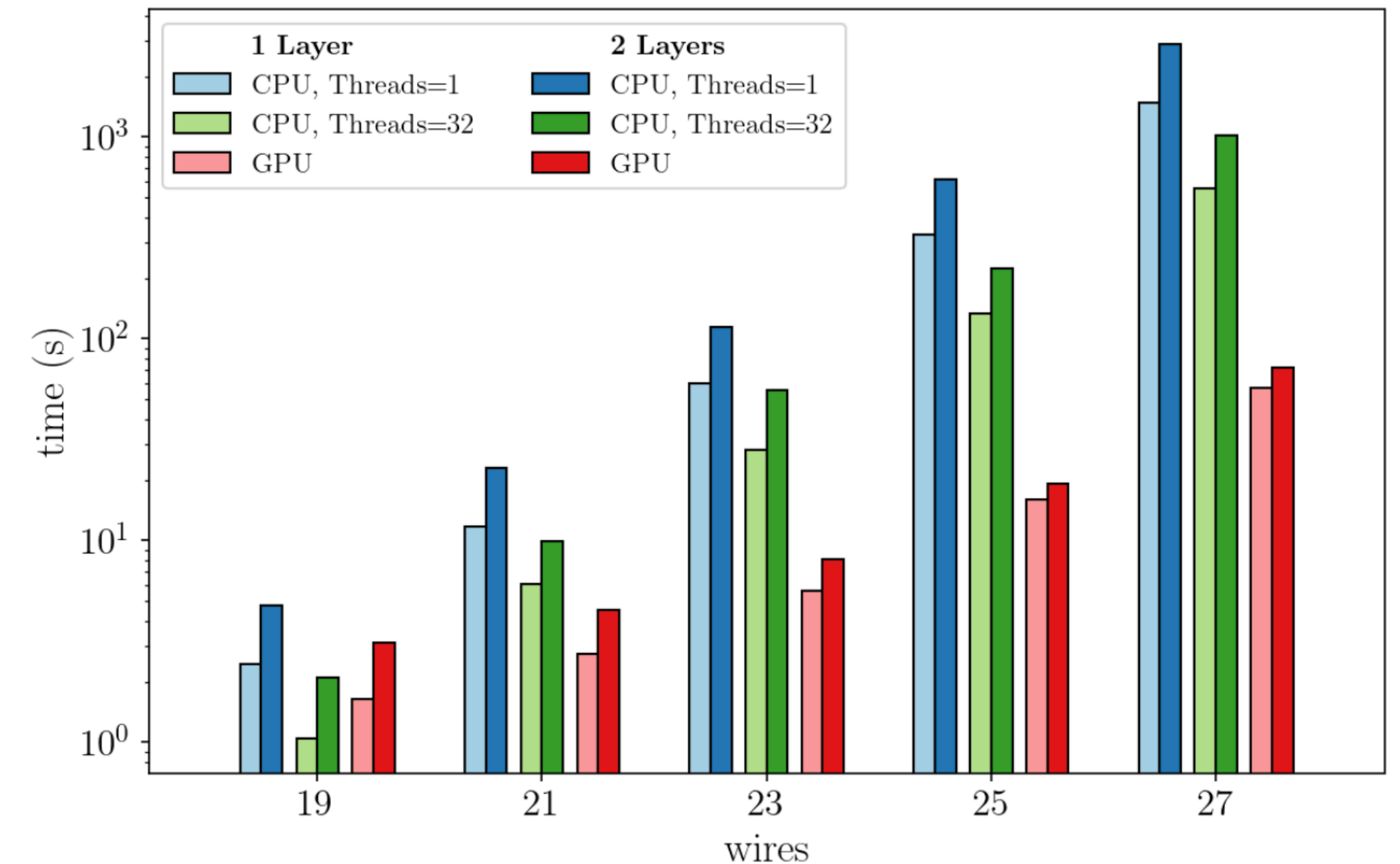
RECAP FOR QR4EO PROJECT

- We evaluate the performance of the hybrid model by the Mean Absolute Error (MAE) and the Continuous Ranked Probability Score (CRPS) on the 10 meter wind fields u_{10m} (west-east) and v_{10m} (south-north)
- The hybrid model results shows comparative performance compares to classical model in the same setting
- But the tested circuit size is limited to 12-qubit and up to 3 circuits was simulated, corresponds to $9/128$ latent channels



FROM HYBRID DIFFUSION POC TO SCALING

- It would be an interesting research direction to investigate: **how does useful quantum capacity scale before costs dominate?**
- To scale up, there are three main challenges:
 - Computational: statevector cost exponential
 - Architectural: more quantum channels may remove classical capacity and the current ansatz could be efficiently replaced by classical method
 - Quantum Circuit Design: larger circuits are not useful unless added qubits/gates encode useful interaction



Source: pennylane blog

<https://pennylane.ai/blog/2022/07/lightning-fast-simulations-with-pennylane-and-the-nvidia-cuquantum-sdk/>

OUR SCALING PROPOSAL

Simulation scaling

Distributed GPU
statevector
simulation

Hybrid architecture

Capacity-
preserving
quantum insertion

Quantum Feature

Feature-space
contribution
+ Functional
contribution

OUR SCALING PROPOSAL

Simulation scaling

Distributed GPU
statevector
simulation

Hybrid architecture

Capacity-
preserving
quantum insertion

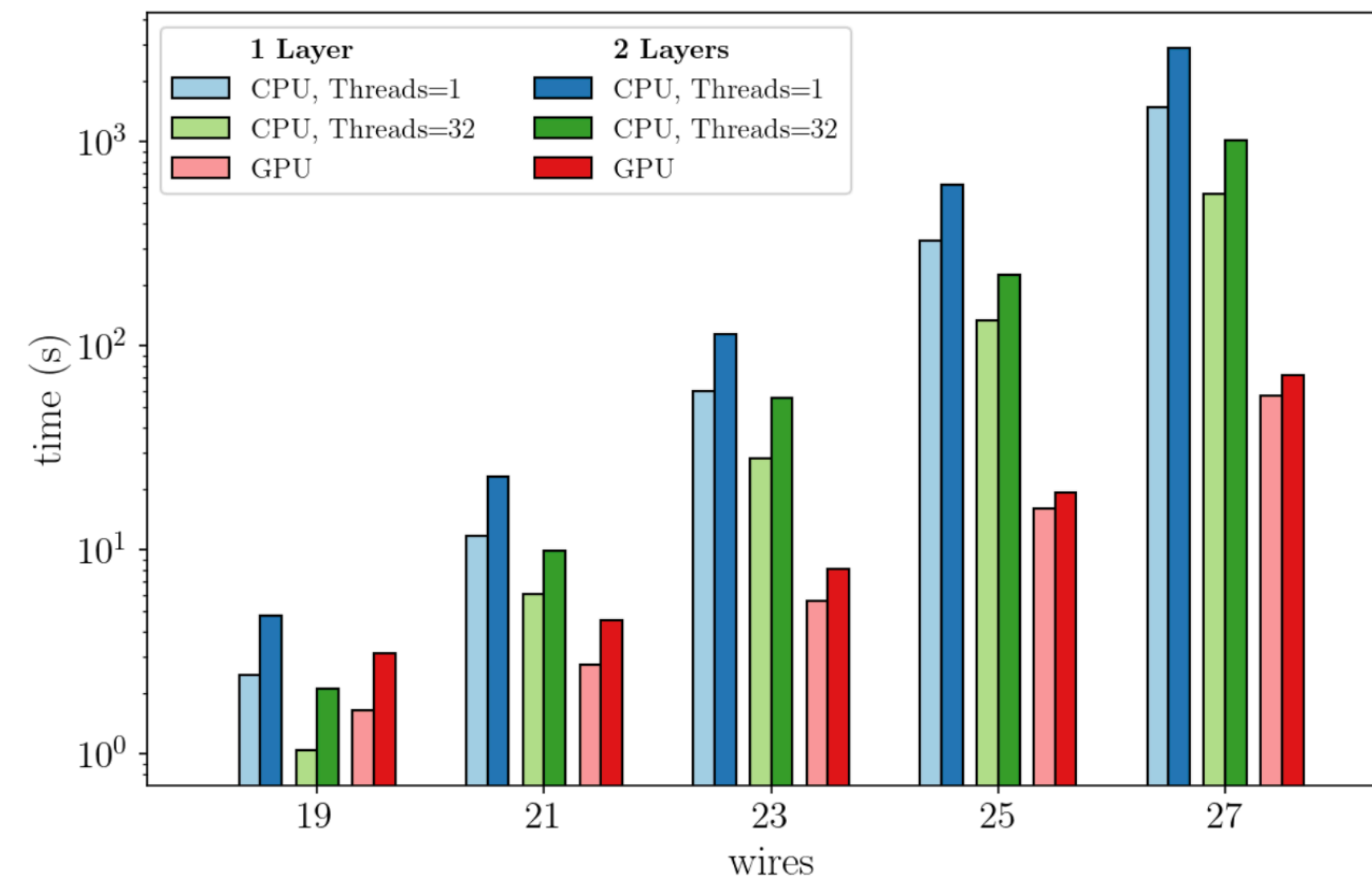
Quantum Feature

Feature-space
contribution
+ Functional
contribution

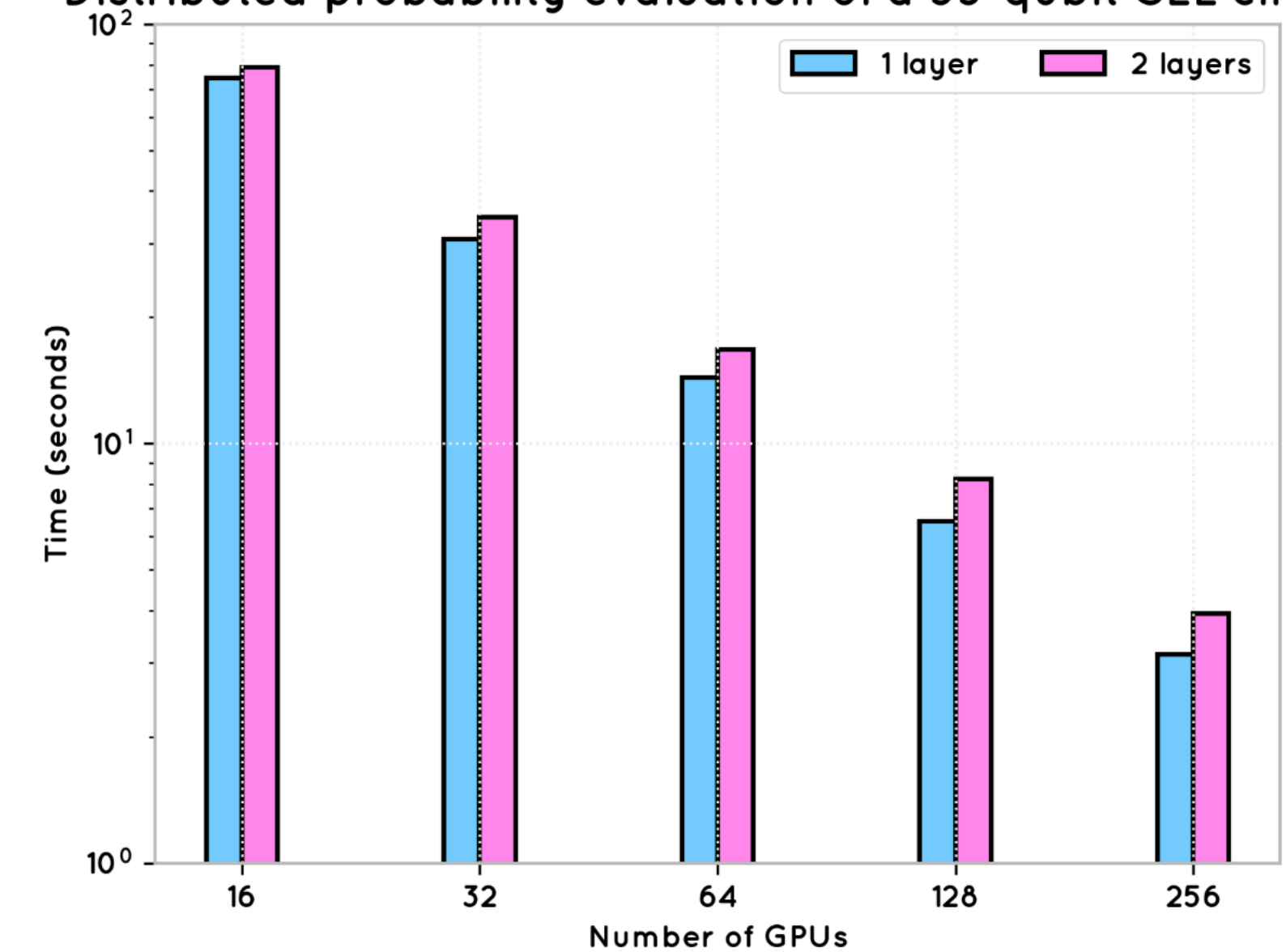
CAN LARGE QUANTUM CIRCUITS BE SIMULATED?

Simulation scaling

- Prior work of PennyLane has shown that large statevector simulations can be extended beyond single-GPU memory limits by distributing one quantum state across multiple GPUs/nodes using lightning.gpu, MPI, and NVIDIA cuQuantum.
- Their demonstrations include >30-qubit workloads, fast distributed adjoint gradients



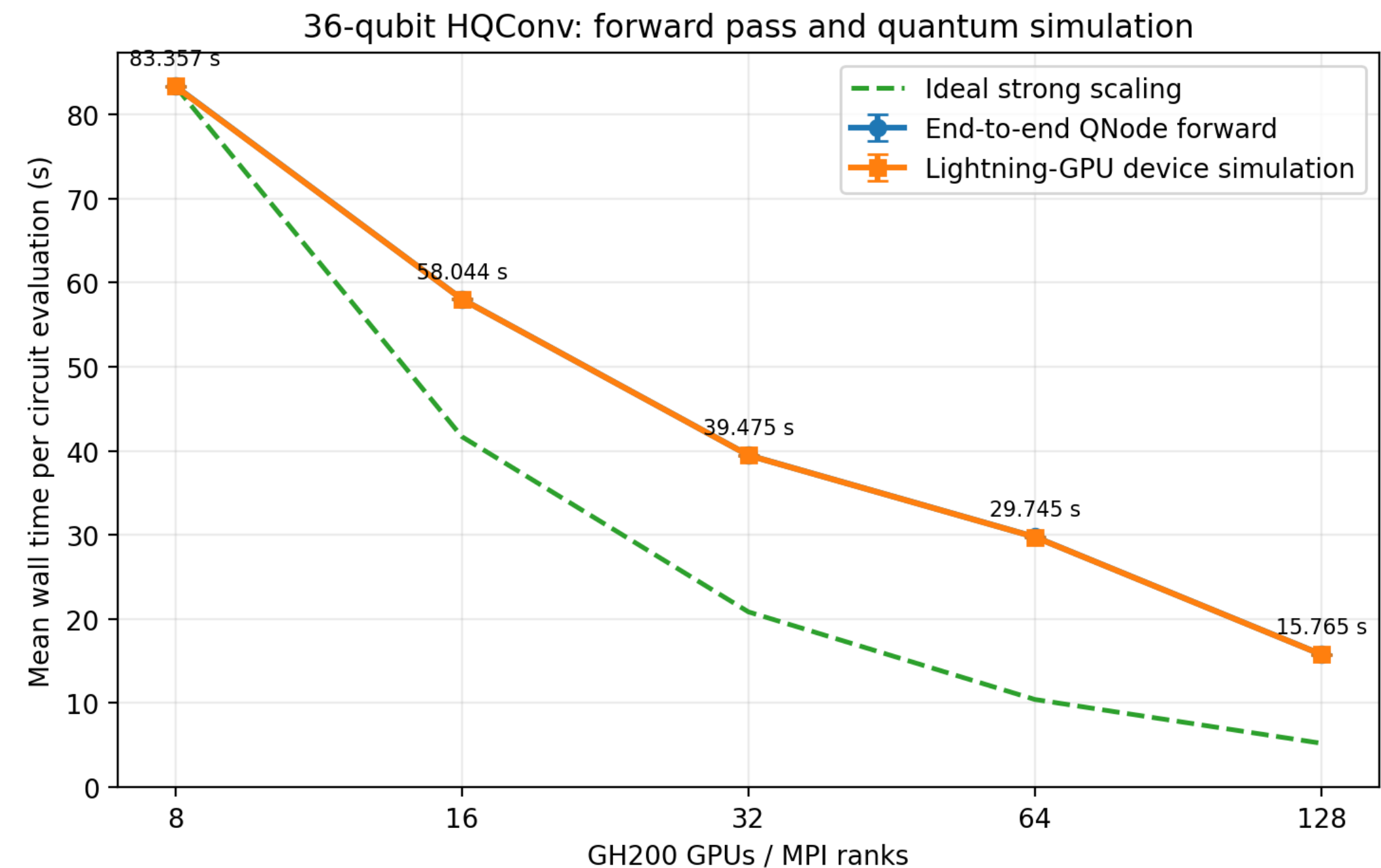
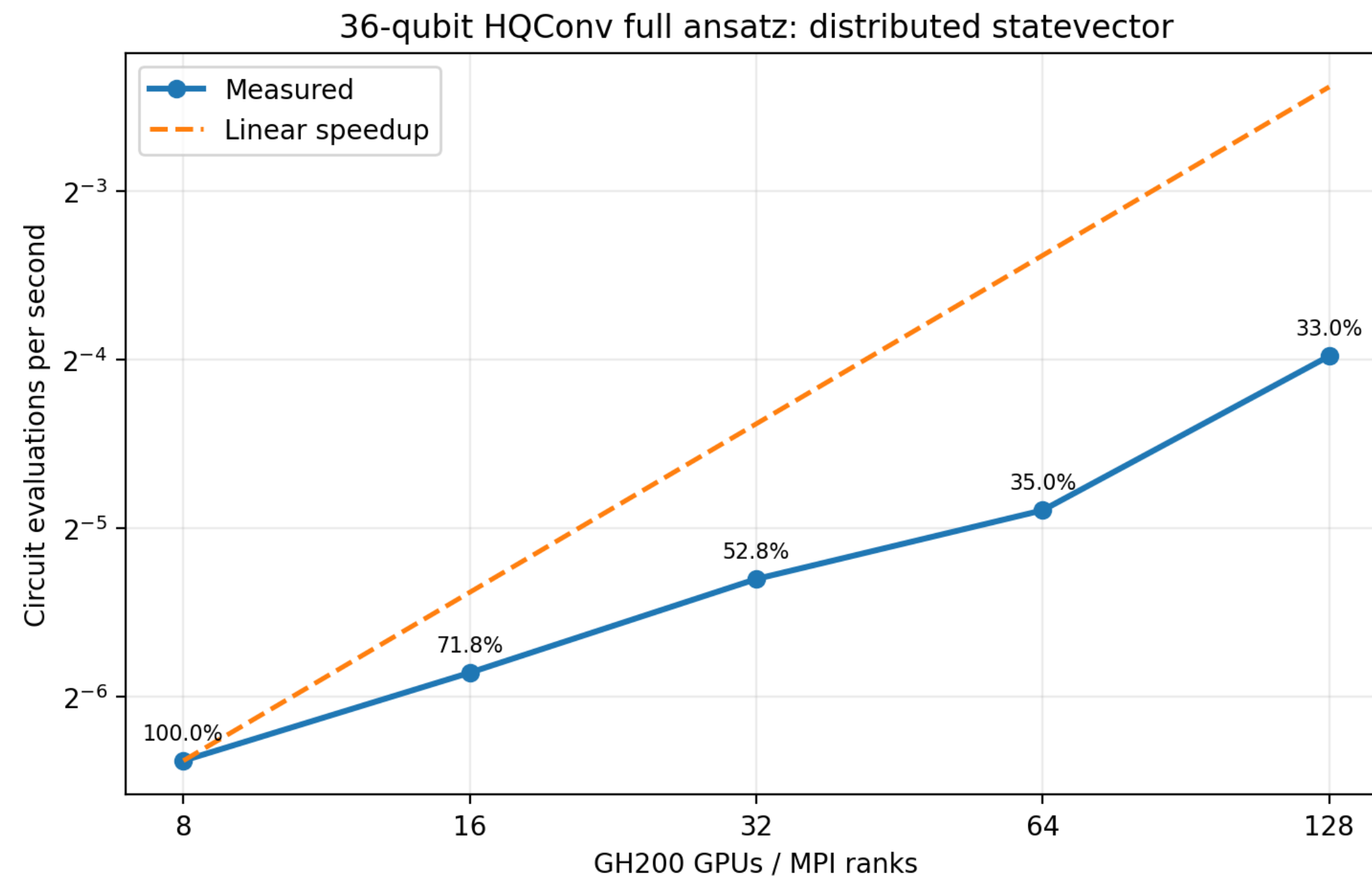
Distributed probability evaluation of a 33-qubit SEL circuit



SCALING OUR TARGET CIRCUIT ON JUPITER

Simulation scaling

- Motivated by this distributed-simulation paradigm, we test whether the quantum circuits required by our hybrid diffusion model exhibit similar scaling on JUPITER



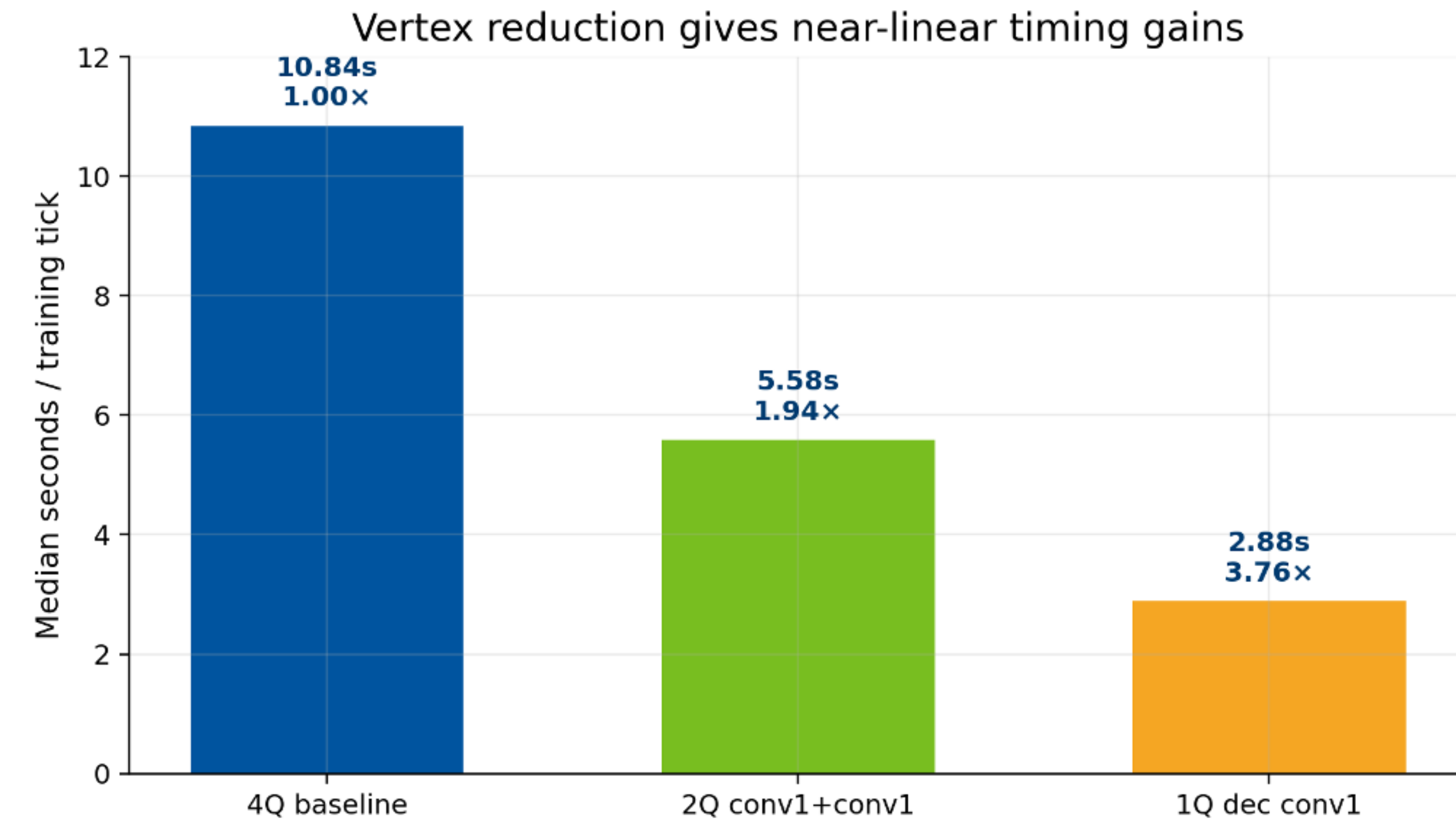
FURTHER SCALING DIRECTIONS

Simulation scaling

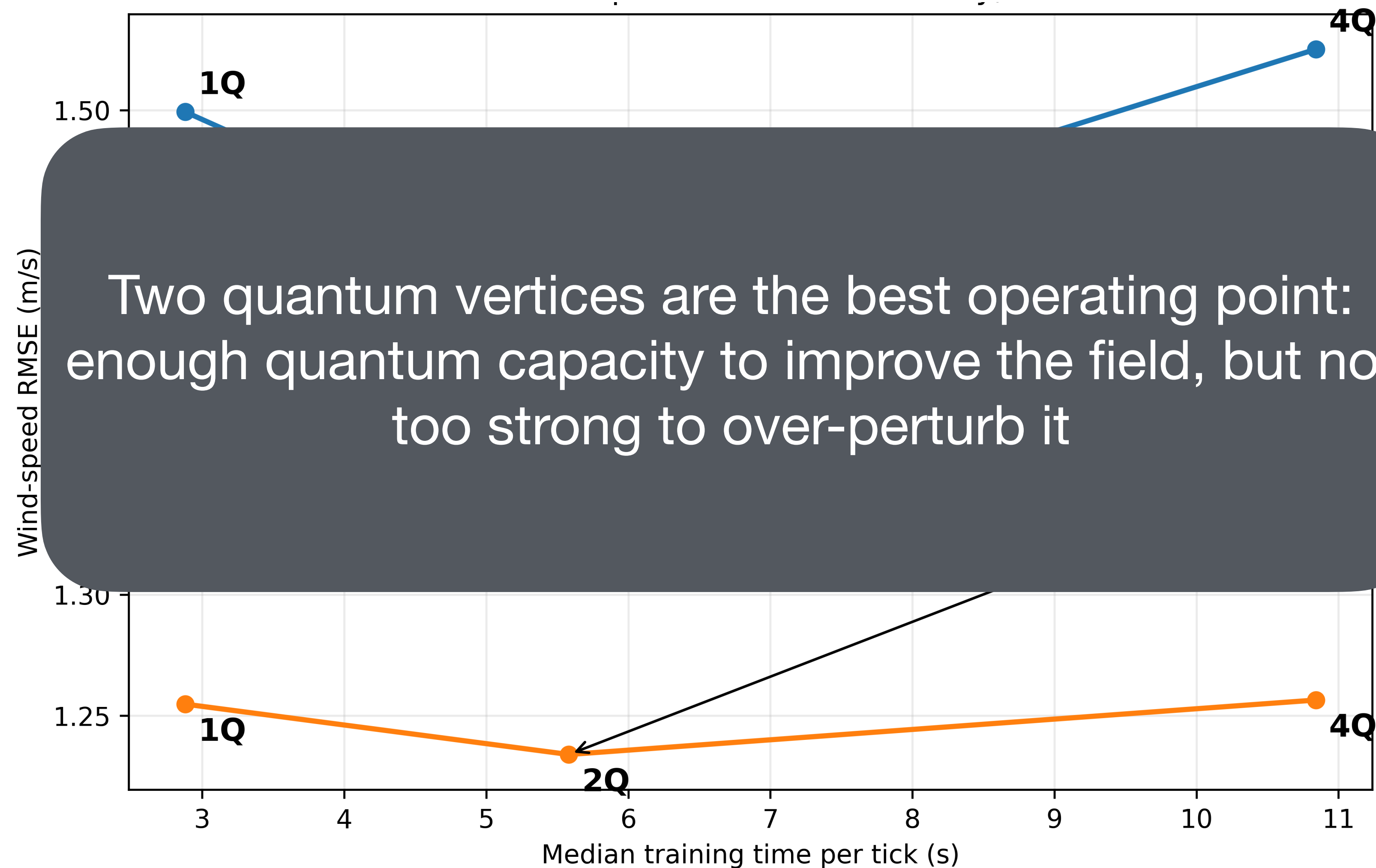
- Multi-node state vector simulation can obtain sub-linear acceleration, but 15.765 second for a forward pass is still not fast enough for a 8 million sample model training:
 - If we assume forward time $\approx$ backward time, a full model training would approx. cost:
15.765s (single pass) * 2 (forward+backward) * 8,000,000 samples * 288 cores * 32 nodes = 726,451,200 core hours
- One possible way of time reduction is reducing the number of quantum insertions: the current version has 4 quantum vertices, is that possible we obtain the same downstream performance with reduced quantum insertions?
- Another direction to explore is to discover which batch size / total trained sample size / backend is the sweet spot of execution speed and downstream performance

VERTEX REDUCTION: QUANTUM VERTICES DOMINATE TRAINING TIME

- 4Q = encoder/decoder conv0+conv1 quantum vertices
- 2Q = encoder conv1 + decoder conv1 quantum vertices
- 1Q = decoder conv1 only



VERTEX REDUCTION: QUANTUM VERTICES DOMINATE TRAINING TIME



OUR SCALING PROPOSAL

Simulation scaling

Distributed GPU
statevector
simulation

Hybrid architecture

Task-aligned
circuit design

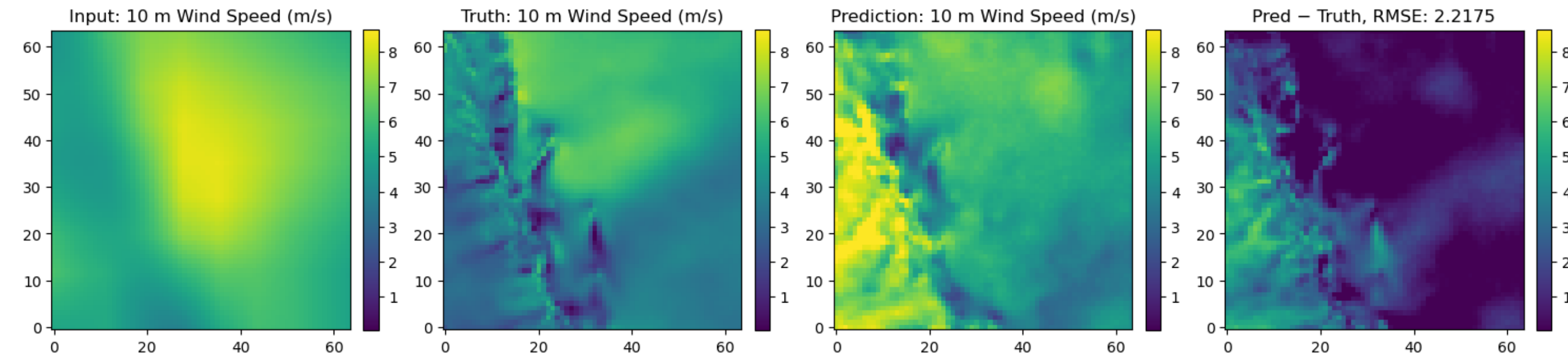
Quantum Feature

Feature-space
contribution
+ Functional
contribution

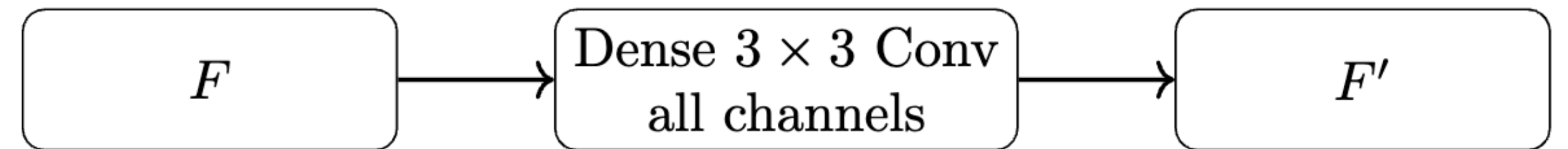
HIDDEN ISSUE IN THE ORIGINAL HYBRID SCALING

Hybrid architecture

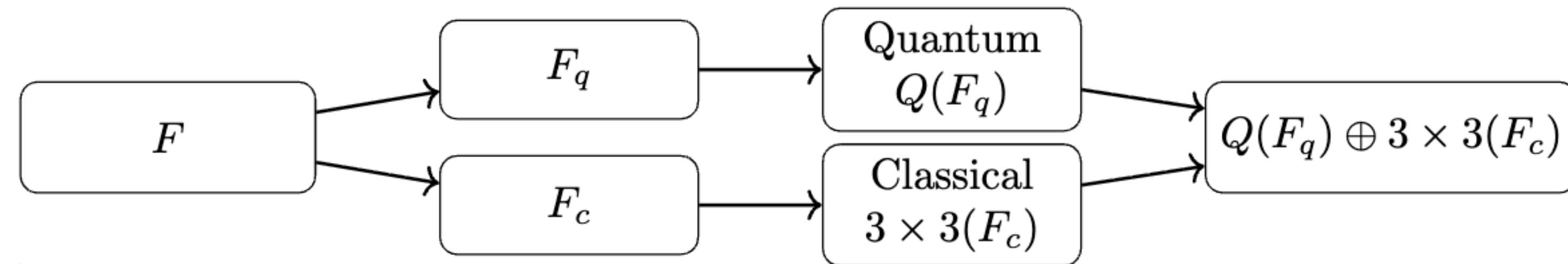
- In the previous experiments, when # of circuits increase to 4 and 5, the generation performance degrades
- This is likely because in the previous hybrid regime, we replace the classical 3x3 convolution channel to $Q(F_q) \oplus 3 \times 3(F_c)$, which implies that more quantum channels leads to less classical mixing
- Previous circuit-count scaling simultaneously changed quantum capacity **and** classical capacity



Native classical block



Replacement hybrid

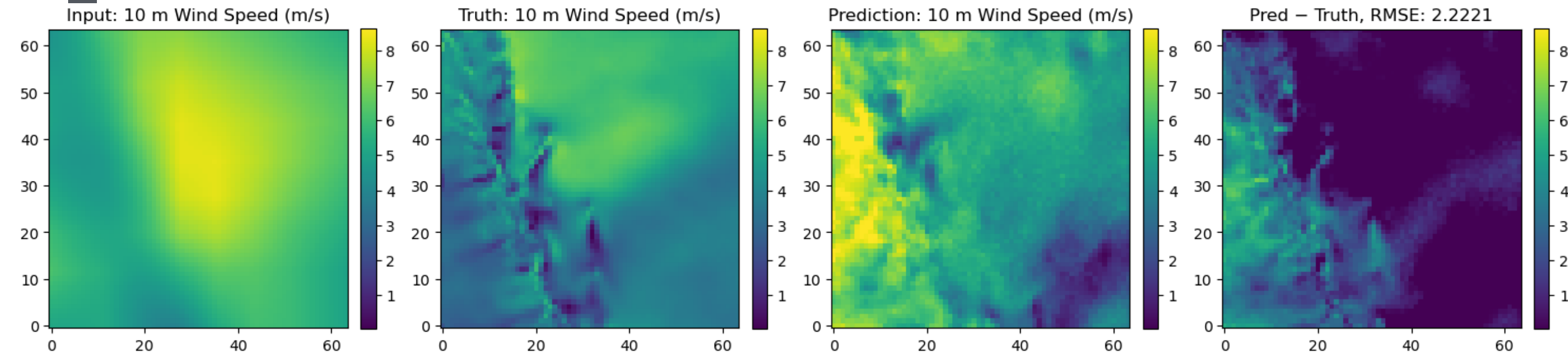


TESTING THE CAPACITY-LOSS HYPOTHESES

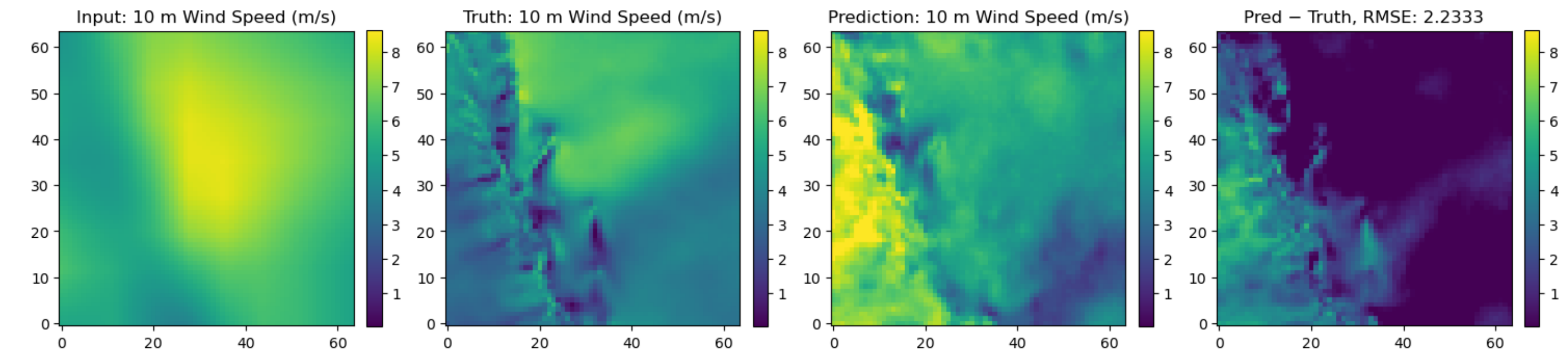
Hybrid architecture

- We did a probe experiment that replace each quantum branch with a classical branch of matching input/output width
- With the increase of replacement ratio, the downstream RMSE increases

n_circ=4



n_circ=5

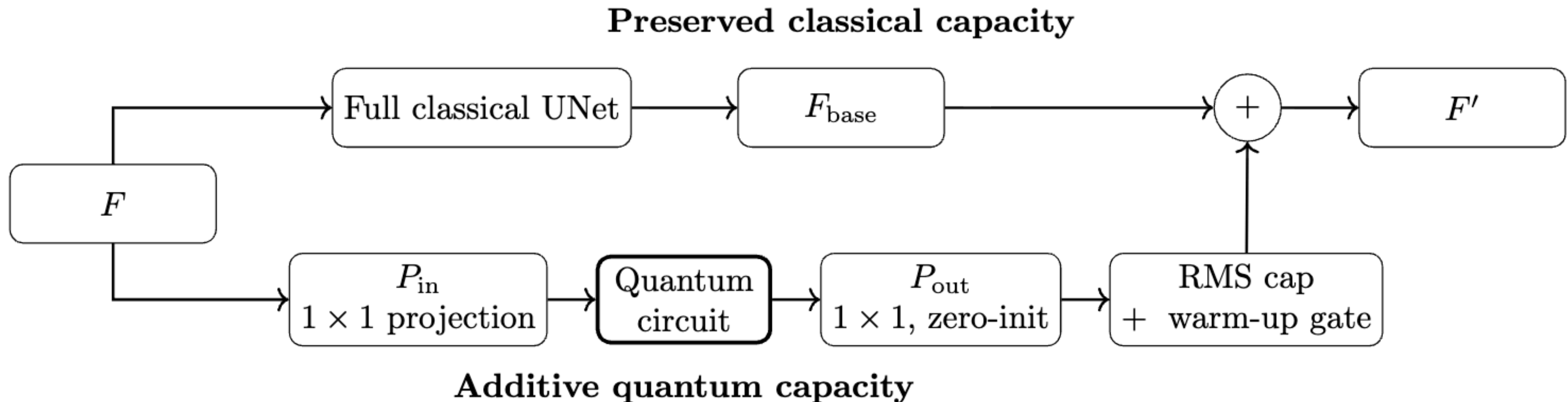


CAPACITY-PRESERVING HYBRID SCALING

Hybrid architecture

- Instead of direct replacement, we introduce an extended way of adding the quantum capacity:

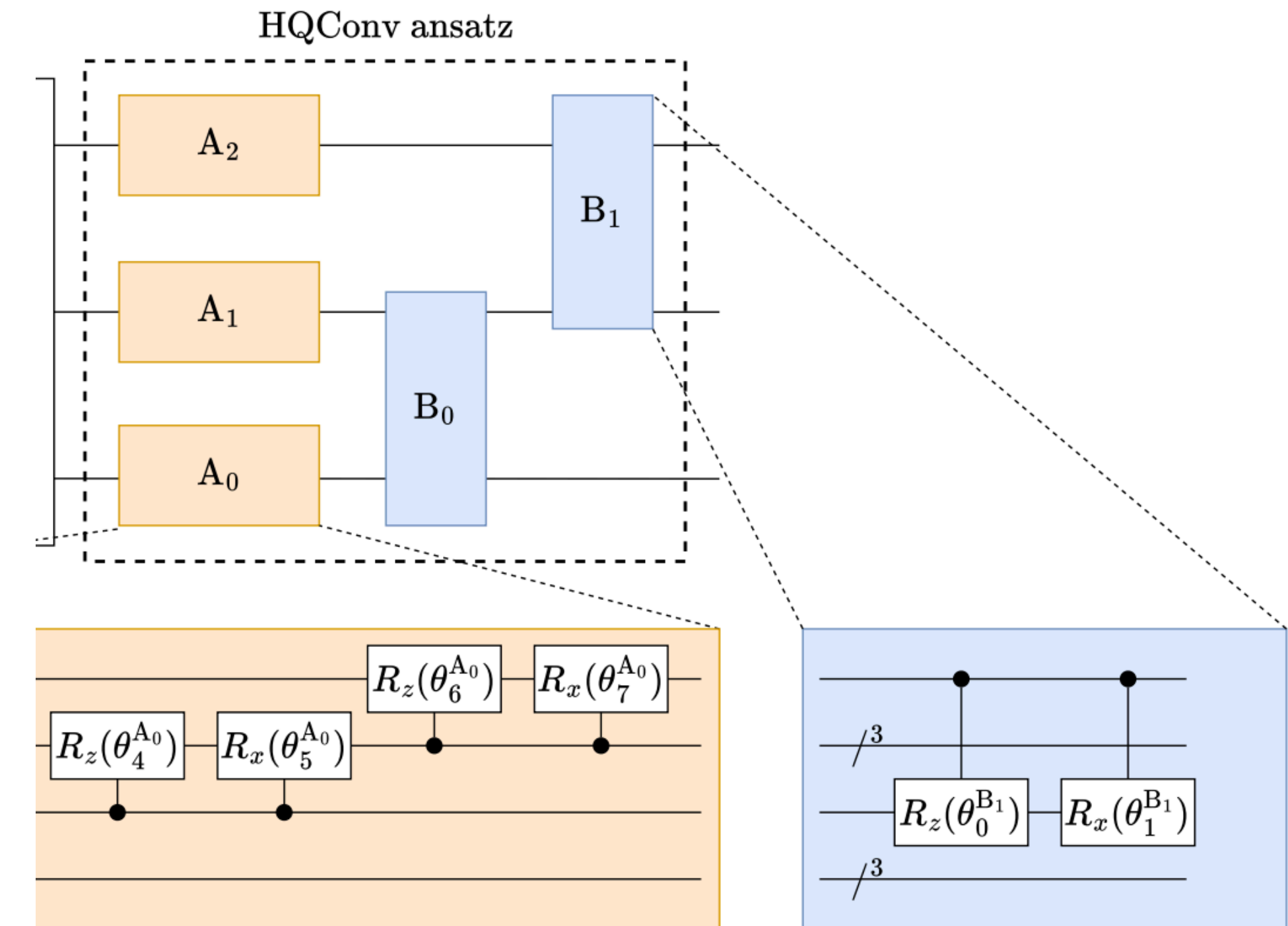
$$F' = F_{\text{base}} + g \text{Cap} \left[P_{\text{out}} Q(P_{\text{in}} F) \right]$$



DRAWBACK OF THE INTERCHANNEL INTERACTION

Hybrid architecture

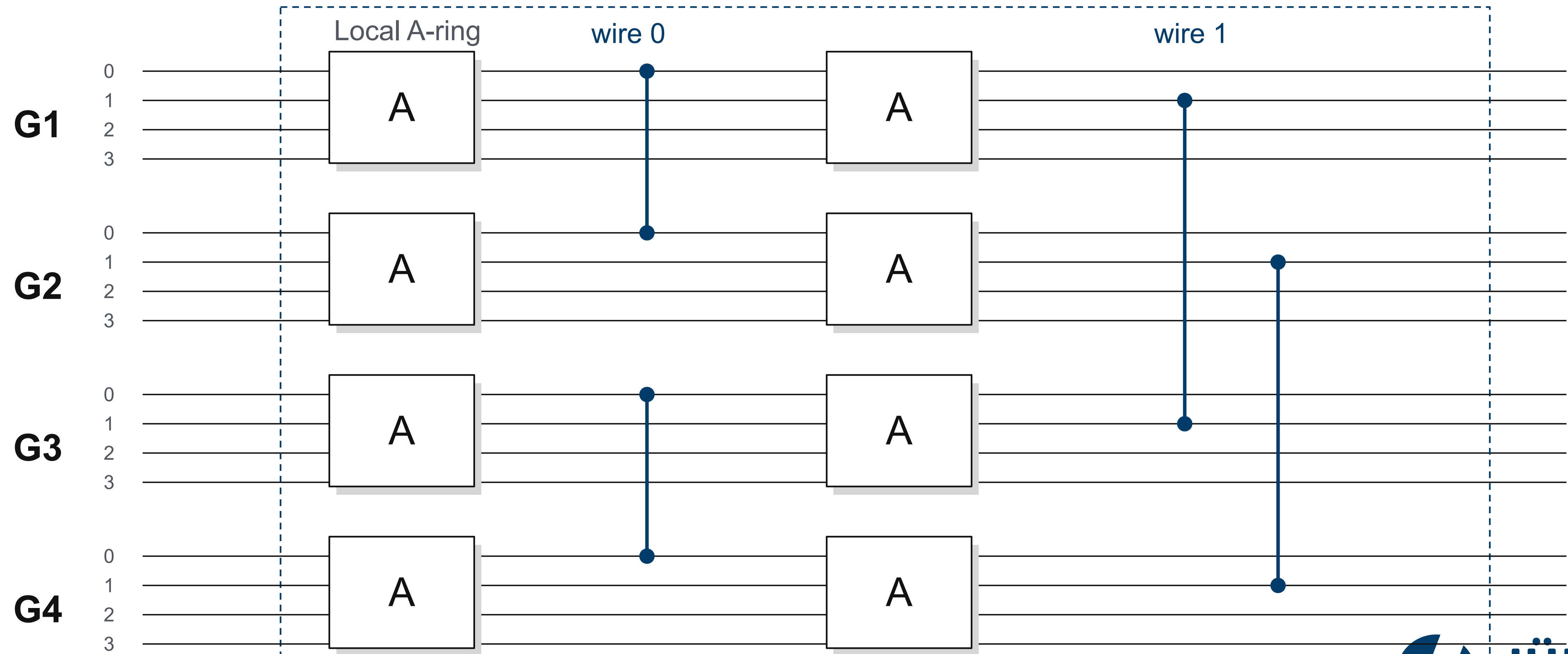
- HQConv B-block = one gateway qubit per channel + cross-channel coupling among gateways
- Reasonable when we only have 3 channels, but only the first qubit is responsible for the communication
- Each channel has many local features, but only one bottleneck token is permitted to perform cross-channel coupling



BETTER (MAYBE) DESIGN: BUTTERFLY STYLE

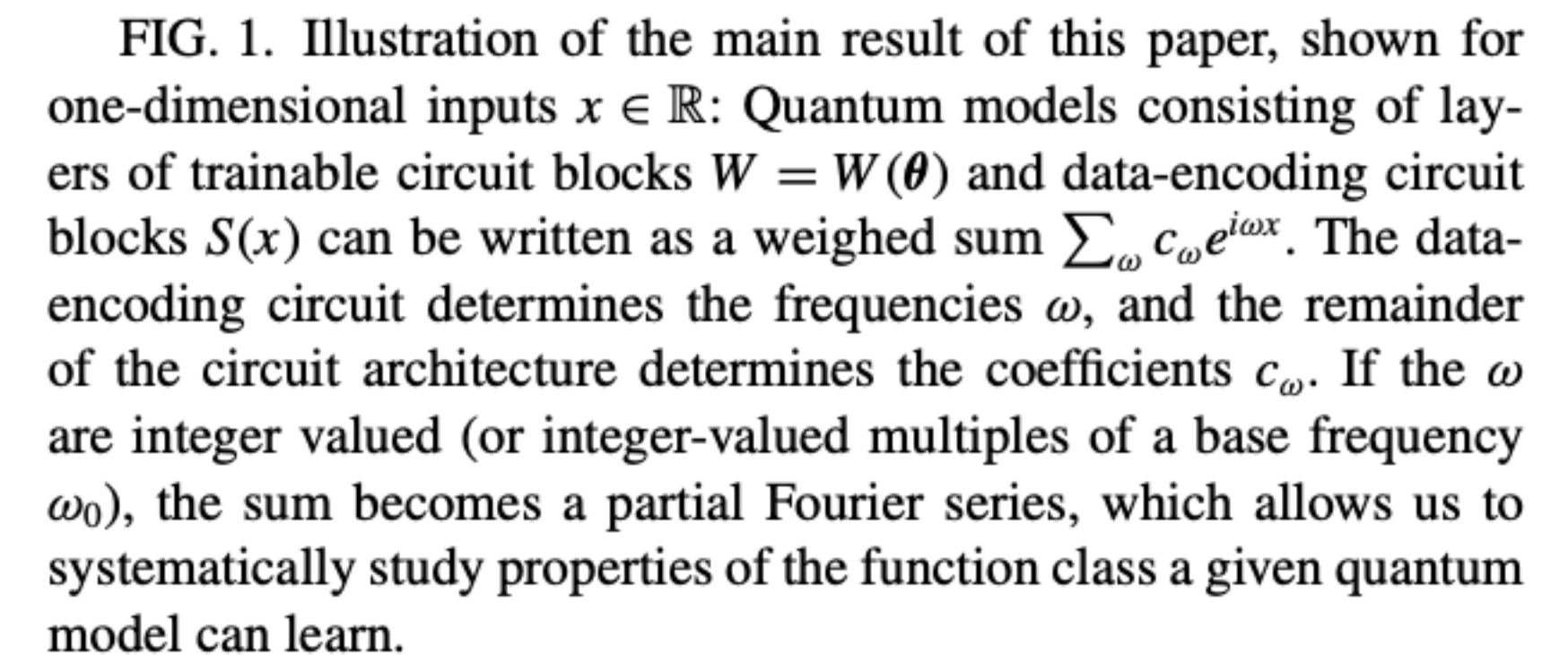
Hybrid architecture

- Inspired by the butterfly network, we design a butterfly style interchannel connection



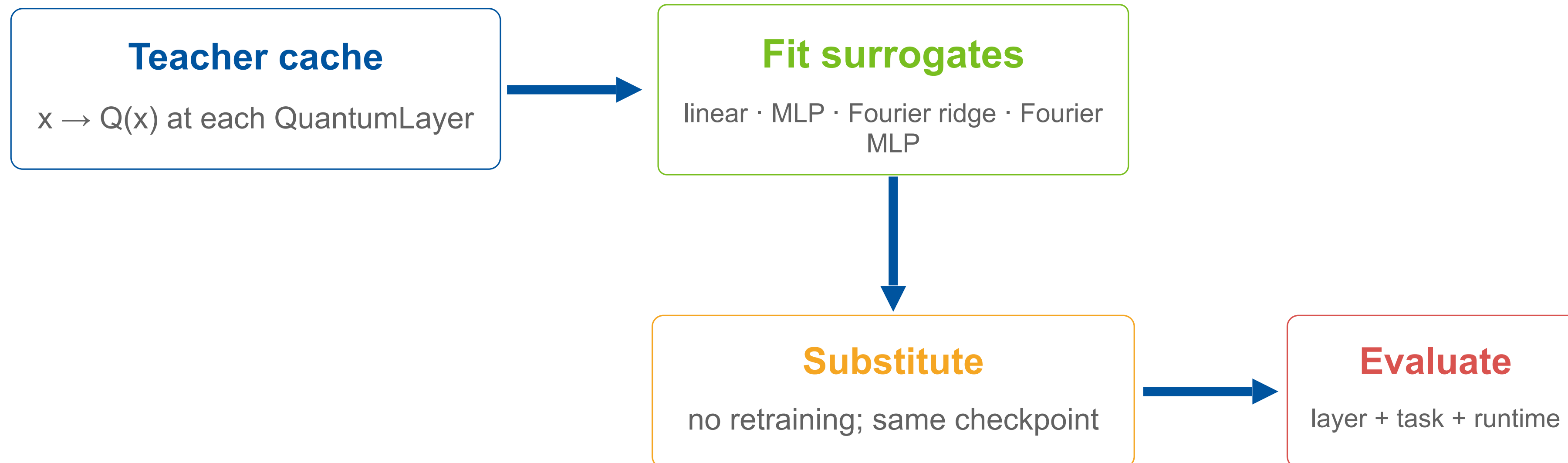
Hybrid architecture

- This motivates that it is possible to replace the VQC with a cheaper classical kernel



DEQUANTAZATION OF THE HQCONV

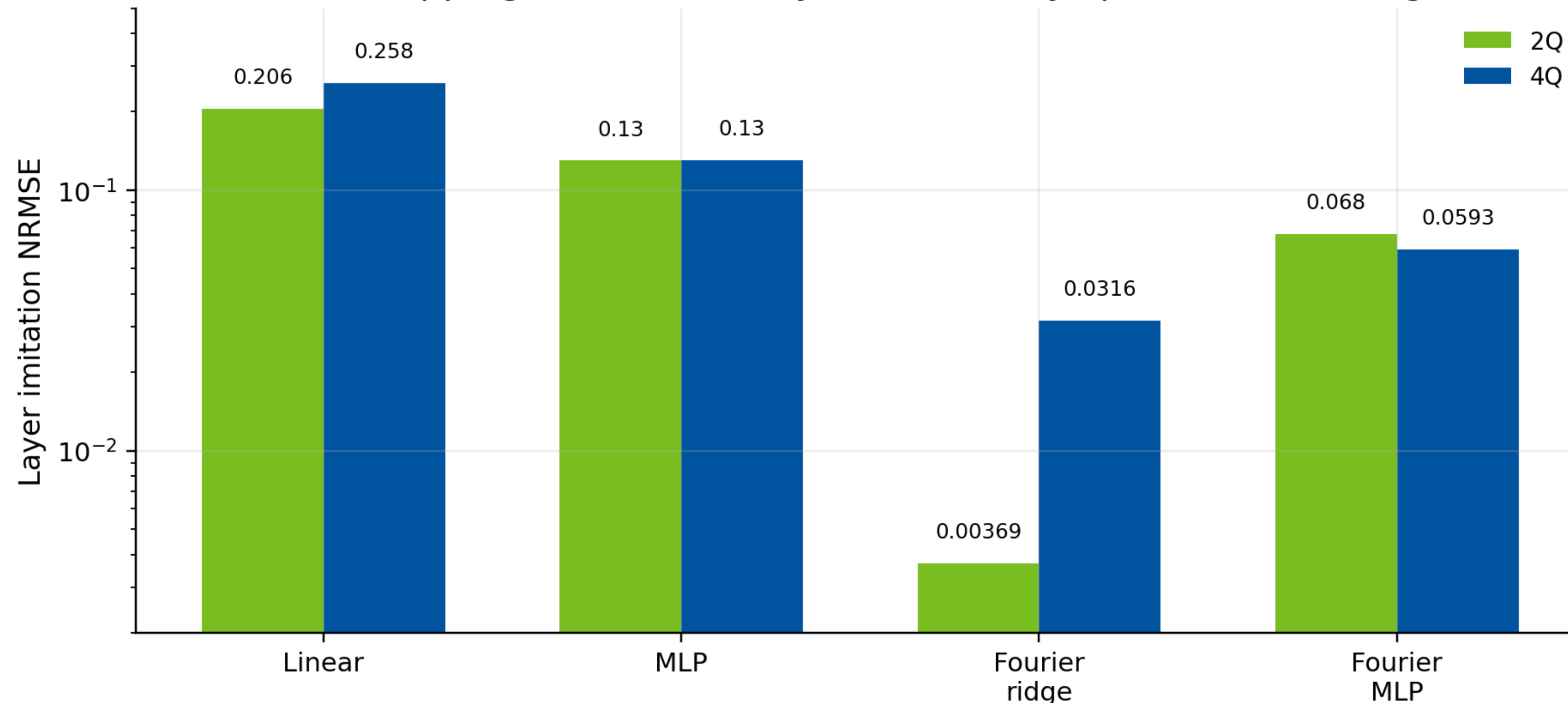
Hybrid architecture



FUNCTION-LEVEL RESULT

Hybrid architecture

PQC mappings are accurately recovered by sparse Fourier ridge



- Linear is insufficient: HQConv is not merely affine on the latent manifold
- Generic MLP helps, but Fourier ridge is best
- This supports functional dequantization of current 12q HQConv

ANSATZ SEARCH

Hybrid architecture

Stage A — structural search at 12q / 16q

- Use trainability / complexity diagnostics from Peter Röseler (@QIP group in JSC)

- Add custom **Cross Group Jacobian**:

$$S_{\text{cross}} = \frac{\| \partial q / \partial x \|_{\text{off-group}}}{\| \partial q / \partial x \|_F + \epsilon}$$

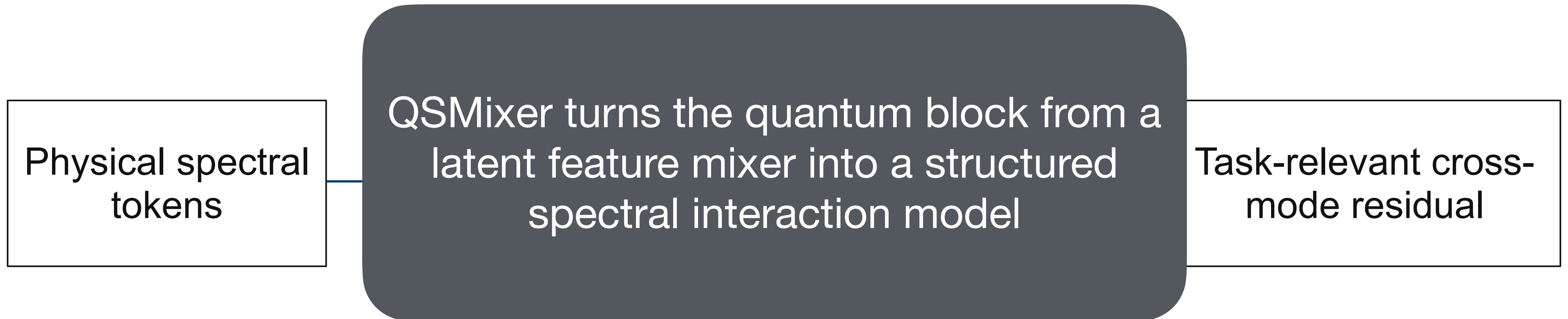
- Retain candidates with:
 - usable gradient variance
 - strong cross-group influence
 - moderate depth / 2Q-gate cost

Stage B — dequantization-aware screening

- Re-evaluate only top 10–20 candidates
- Fourier-ridge surrogate budgets: 16, 32, 64, 128, 256, 512
- Measure $C_F(\epsilon)$: minimum Fourier-feature budget required to approximate PQC
- Also record:
 - forward + gradient runtime
 - circuit depth / gate count

NEXT ANSATZ DESIGN:QSMIXER

Hybrid architecture



Input-conditioned interaction across selected modes

OUR SCALING PROPOSAL

Simulation scaling

Distributed GPU
statevector
simulation

Hybrid architecture

Capacity-
preserving
quantum insertion

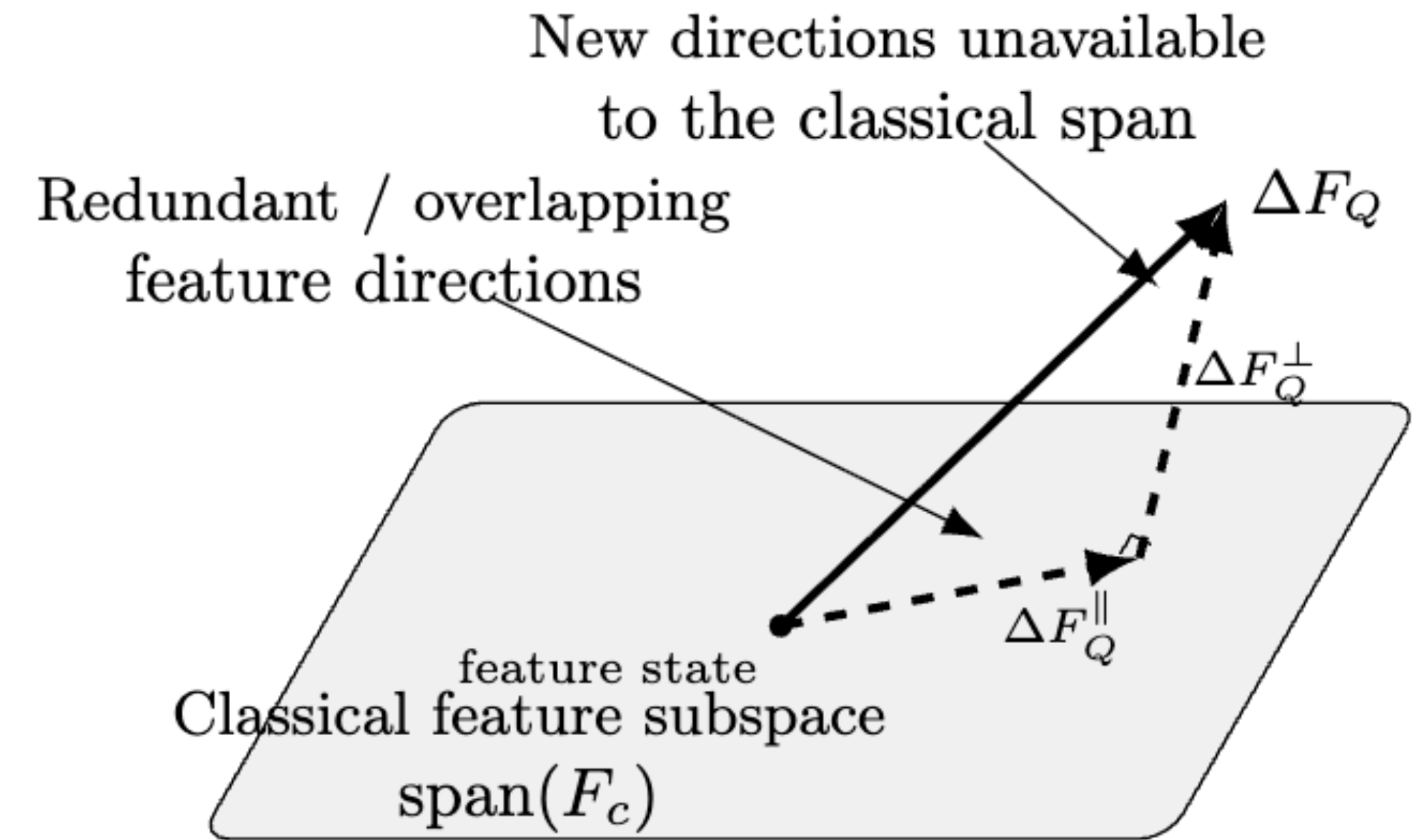
Quantum Feature

Feature-space
contribution
+ Functional
contribution

FEATURE-SPACE CONTRIBUTION

Efficient quantum capacity

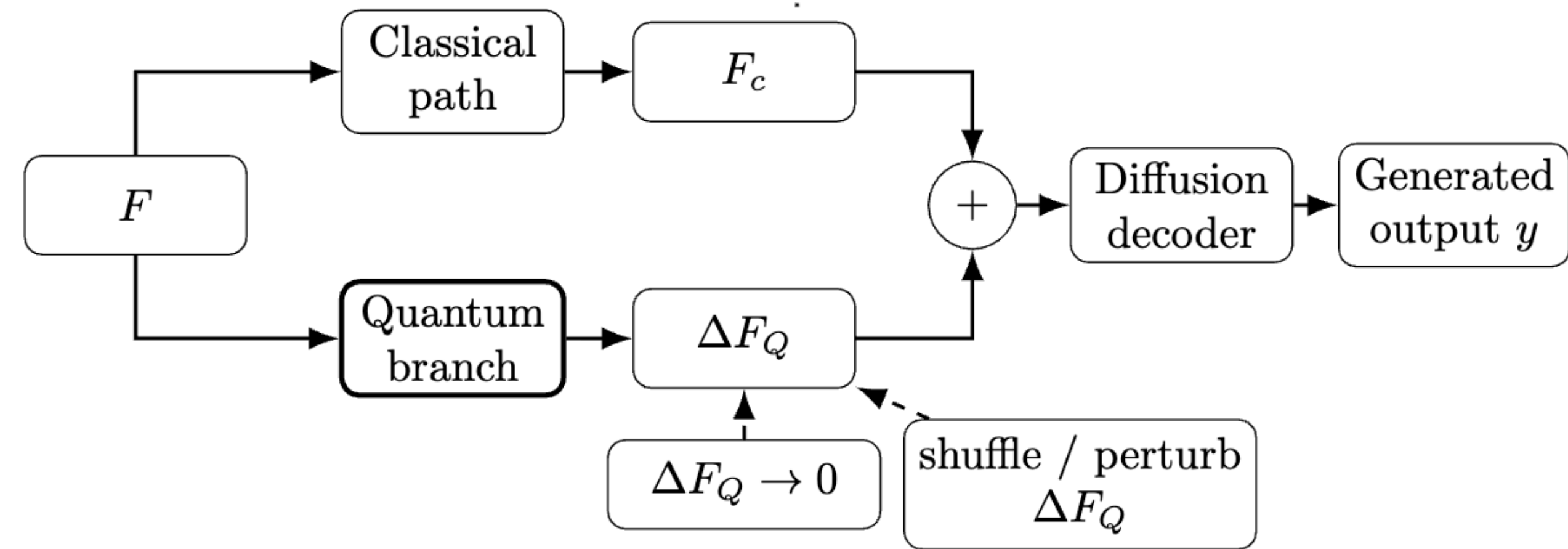
- We want to answer the fundamental question: does the quantum branch add new feature directions?
- It is possible that the quantum features are:
$$\Delta F_Q = \Delta F_Q^{\parallel} + \Delta F_Q^{\perp}$$
- Useful quantum scaling should increase non-redundant feature directions
- Possible diagnostics to test the direction:
 - subspace similarity: CKA / SVCCA / principal angles
 - effective rank / covariance / singular-value spectrum



FUNCTIONAL CONTRIBUTION

Efficient quantum capacity

- We also want to answer: does the diffusion model actually use these features?
- Possible ways to test:
 - Branch ablation: whether removing the quantum branch degrade RMSE / CRPS / spectra
 - Shuffle or perturb quantum branch across samples.



Functional intervention on the trained model

ROADMAP

Stage I — Identify the optimized quantum runtime

Stage I — Check whether the architecture is the optimal for scaling

Stage III — Determine useful quantum capacity

Stage IV — Scale qubits / nodes: $12q \rightarrow 28q \rightarrow 32/36q; 1 \rightarrow N$
GH200 nodes