UTILIZATION OVER QUBITS: RYDBERG REGISTERS IN PHYSICS WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In hybrid classical–quantum models, how many qubits are available matters less than how well they are used. We attach an input-dependent Rydberg-atom co-processor, QUOTA, to world models on ten industry benchmarks (fusion, geothermal, fluid flow and vehicle routing; mostly released state-of-the-art models), under pre-registered rules. With the same 30 qubits, it cuts error by up to 36%, against 7% for the standard design that averages features before the qubits see them. With the classical model fixed, error falls steadily as the budget grows from 0 to 50 qubits, by 13% on shear flow and 6% on convection, following a closed-form hybrid scaling law whose 50-qubit prediction, registered before the runs, held on every seed. The gains come from allocating by signal type: the error of each prediction is split into frequency bands (Fourier octaves and wavelet tiles), the smooth, predictable (analog) bands go to the qubits, the rest (digital) stays with the classical network. For each input, the atom positions, detunings and tethers between atom groups are also set anew. Drawing the per-input configuration from a noise code, as in functional generative networks, turns the hybrid into an ensemble whose energy score is 16–20% lower. Most of this configuration maps onto current neutral-atom processors such as Pasqal's, one device program per input, once the detunings are re-encoded to fit the device's local addressing. For a fixed classical model, the hybrid scaling law reads $G(K) = \eta U_\infty[1-(1+K/K_0)^{1-a}]$, the share of error energy removed by K qubits, where U_∞ is the share the qubits can reach, η how well they use it and $a \approx 2$; it tells how many qubits a task can put to use.

1 INTRODUCTION

Learned physics simulators are now central to scientific machine learning, from turbulence and magnetohydrodynamics (Ohana et al., 2024; McCabe et al., 2025) to tokamak plasmas (Boschi et al., 2026; Rousseau et al., 2026) and real-world flow measurements (Hu et al., 2026). In parallel, a fast-growing literature inserts small quantum circuits into such models: hybrid neural operators (Jain & Kalev, 2026; Hossain et al., 2026; Marcandelli et al., 2025), quantum adapters for frozen networks (Koike-Akino et al., 2025; Liu et al., 2025; Raj & Coyle, 2025), and quantum-informed priors for turbulence (Wang et al., 2026a; Xue et al., 2026). Whether the quantum part does anything that a classical module with the same interface and parameter count could not do is rarely tested; where it has been, apparent edges tend to evaporate under equal-tuning classical baselines (Bowles et al., 2024; Pandey, 2026a).

We run that test at scale, on the analog neutral-atom model family that current hardware natively evolves (Kornjača et al., 2024). QUOTA (*Qubit Utilization thrOugh Trained Allocation*) is a Rydberg register (one atom is one qubit) whose allocation is learned, not hand-placed. Per-atom detunings are functions of each input; geometry and drives are trained by gradient; register size and graph are chosen by architecture search; band-limited or routed insertions decide which modes (or, per input, which error tiles) the atoms serve. It is inserted into the *released* checkpoints of published models and compared with a parameter-matched MLP core behind the identical interface, a no-interaction ablation, and a mean-field control, under kill rules fixed before each run. The premise is a division of labor: a pretrained classical model handles the easily digitized part of a field, and a small qubit budget is worth only as much as it is utilized on the part each input gets wrong. We separate two claims. Utilization, not qubit count, governs the gain. Whether a quantum core exploits

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

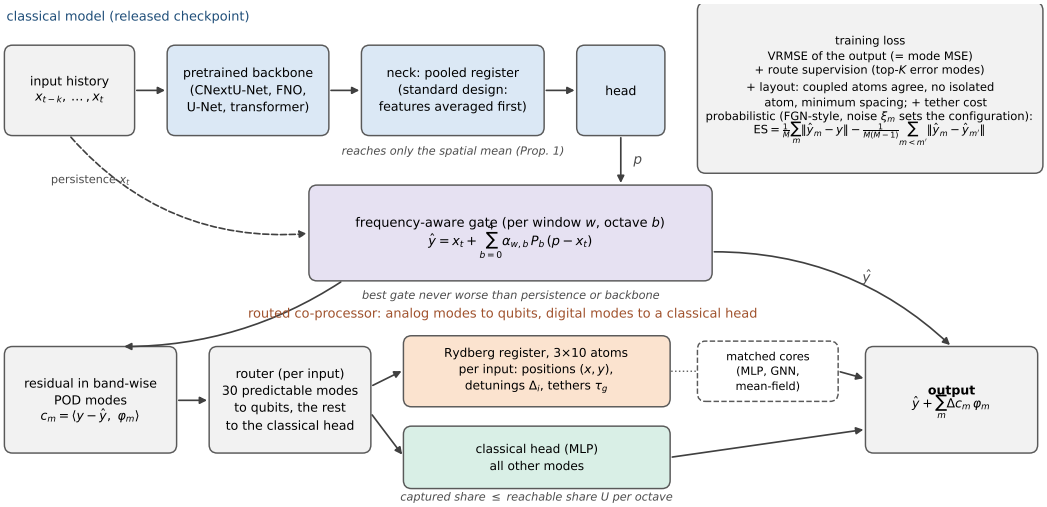

Figure 1: The hybrid. Top: the released classical model (backbone, neck, head), with the standard pooled register in its neck. Middle: a frequency-aware gate mixes backbone and persistence per window and octave (P_b: projection onto octave b). Bottom: the remaining error, in band-wise POD modes, is routed per input: 30 predictable modes to a Rydberg register whose positions, detunings and tethers depend on the input, the rest to a classical head; matched classical cores swap into the register's slot. Italics: bounds proved in Lean 4.

a given utilization better than a classical one is a second claim; at the sizes we can simulate, matched classical cores do as well.

Contributions.

1. **Input-dependent allocation of qubits** (§3, §5.5, Fig. 1): each prediction's error is split into Fourier octaves and band-wise POD modes; the smooth, predictable (analog) modes go to 30 Rydberg atoms whose positions, detunings and tethers are set per input, the rest (digital) to a classical head, behind a frequency-aware gate between backbone and persistence. With the same 30 atoms error falls by 0.9–36% across rows, 36% on convection against 7% pooled.
2. **A hybrid scaling law** (§5.5, Appendix N): with the classical model fixed, error falls steadily from 0 to 50 qubits, by 13% on shear flow and 6% on convection, following a closed form whose 50-qubit prediction, registered before the runs, held on all six seeds.
3. **Proofs of where qubits can help** (§5.3, Lean 4): a register fed spatially averaged (*pooled*) features, the standard design, is blind to every zero-mean mode, so their reach is the DC share of the error (Proposition 1); routed corrections are bounded per Fourier band; the optimal band gate is never worse than persistence or the backbone.
4. **A matched-control audit on released models** (§5.1): eight physical benchmarks (turbulence, MHD, convection, shear flow, tokamak transport and equilibrium, geothermal flow), real-world flows (RealPDEBench) and vehicle routing with two released solvers (Appendix P), with parameter-matched classical cores (MLP, GNN on the same learned graph, mean-field register) in the register's slot. The matched cores capture the allocation gains as well, so the gains belong to allocation; a long-horizon TRL2D edge (§5.2) did not replicate on a registered fourth seed.

2 Related work

Quantum modules in classical surrogates. Hybrid FNO and U-Net variants (Jain & Kalev, 2026; Hossain et al., 2026; Marcandelli et al., 2025) and quantum-informed turbulence priors (Wang

et al., 2026a; Xue et al., 2026) typically report gains without parameter-matched cores or released-checkpoint anchors; the closest in-model study, a 5-qubit mixer in a 3-D FNO, beat a larger dense mixer but not its matched MLP-mixer (Papierz et al., 2026). Quantum adapters for frozen language models (Koike-Akino et al., 2025; Liu et al., 2025) use quantum circuits to *generate* classical weights, so inference is classical by construction; our inference-inertness finding reaches the same deployment regime empirically, the classical-surrogate regime of Schreiber et al. (2023). Reservoir and attribution studies show that apparent quantum edges shrink under equal-tuning classical baselines (Pandey, 2026a;b) and reach nulls with matched controls and multiplicity correction (Chen, 2026; Gillani et al., 2026; Legziel Levy et al., 2026; Herbort et al., 2026); the register-free QSWM finds a short-horizon edge on cellular automata (Jiang et al., 2026). We follow the benchmarking cautions of Bowles et al. (2024) and Schuld & Killoran (2022), adding pre-registration, kill rules, and released-checkpoint anchors across eight domains.

Analog ansätze, dequantization, and scaling. Rydberg reservoirs (Bravo et al., 2022; Kornjača et al., 2024) and quantum neural operators (Wang et al., 2025) report gains against quantum or unmatched baselines. ReCon co-trains atom positions and drives derivative-free (DiBrita et al., 2024), ResQ feeds inputs into detunings on fixed lattices (DiBrita et al., 2025), and PulserDiff differentiates pulses and positions (Abramavicius et al., 2025); position gradients were argued ineffective under the divergent r^{-6} term (de Keijzer et al., 2025), which we bound by clamping distances while training geometry by exact statevector autodiff. Variational models that are both trainable and non-dequantizable exist, but on specially constructed learning problems (Gil-Fuster et al., 2025), and trained quantum models admit classical shadows (Keller et al., 2026); our dequantization suite measures this on a trained world model. Saturation in qubit count has been reported in non-parameter-matched sweeps (Kim et al., 2025; Vyskubov et al., 2026; Zabihi et al., 2026) and is predicted by a frozen-block bound (Qi et al., 2026); our ladder adds a 0-qubit arm and a matched core.

Spectral limits of neural surrogates. Neural networks fit low frequencies first (Rahaman et al., 2019), and neural PDE solvers lose long rollouts through high-frequency error (Lippe et al., 2023). We give the mirror-image statement for insertion modules: a pooled module reads and writes only the DC mode, whatever its core (Proposition 1).

3 QUOTA: THE REGISTER, ITS INSERTION, AND THE CONTROLS

Insertion. A backbone produces a latent field $z \in \mathbb{R}^{d\times N}$ over N spatial positions. The standard insertion pools $h = \frac{1}{N}\sum_s z_{:,s}$, applies a core $c(h) \in \mathbb{R}^m$, and writes back a zero-initialized gated broadcast $z_{:,s} \leftarrow z_{:,s} + \alpha \odot U c(h)$ for every s ($\alpha = 0$ at initialization, so the inserted model starts exactly at the released checkpoint). Six audit rows insert at a field-valued latent (a U-Net or operator neck); two (the Sophelio starter model and TokaMind) insert at a vector latent, where the core sees the whole latent and no pooling occurs. §5.4 replaces this insertion by band-limited and routed ones.

Register core. Each 10-atom register evolves under $H/\hbar = \sum_i \frac{\Omega_\ell}{2} X_i - \sum_i \delta_{\ell,i} n_i + \sum_{i<j} C_6 |r_i - r_j|^{-6} n_i n_j$ with input-conditioned detunings $\delta_{\ell,i} = b_{\ell,i} + \pi \tanh(W_\ell h)_i$ (W zero-initialized), trainable drives Ω_ℓ, times τ_ℓ and 2-D atom coordinates r_i (distances clamped at $0.5\,R_b$), L=4 layers of a first-order split (per-atom $R_z(\delta)R_x(\Omega)$, then the diagonal interaction phase), and $\langle Z_i\rangle$ readout (the routed co-processor also reads $\langle X_i\rangle, \langle Y_i\rangle$). Thirty atoms are three blockade-decoupled registers, keeping trainability fixed across the qubit ladder (Cerezo et al., 2025). Training uses an exact complex64 statevector propagator with autograd, cross-checked against CUDA-Q ($\max_i |\Delta\langle Z_i\rangle| \leq 10^{-6}$ on a 12-qubit circuit).

Controls. Every control implements the identical pool→core→residual interface at matched total parameters (the MLP core's width is set from the register's allocated parameter count; within 0.001% at the TRL2D neck). *Matched MLP:* a one-hidden-layer MLP core. *No-interaction:* $V_{ij} \equiv 0$ from step 0 (a product state; removes entanglement and the interaction nonlinearity). *Mean-field (Gutzwiller):* the interaction phase replaced by its Hartree decoupling $\theta_i = \tau \sum_{j\neq i} V_{ij}\langle n_j\rangle$, which keeps the $V_{ij}(r)$ nonlinearity and trainable geometry but forbids entanglement. The ladder removes structure, then interactions, then entanglement alone.

Cost and deployment. Every experiment, from fine-tuning the released backbones to exact register simulation, ran on a single NVIDIA DGX Spark (one GB10 chip, 128 GB unified memory). Exact simulation is the price: one TRL2D fine-tuning step (released CNextU-net, forward, backward and AdamW, batch 4) takes 504 ms for the backbone, 511 ms ($+1.4\%$) with the MLP core, and 595 ms ($+18\%$) with the 30-atom register. Trained neck registers are inference-inert, so deploying them needs no quantum hardware (Appendix M).

Probabilistic variant. Following functional generative networks (Alet et al., 2025), a noise draw ξ_m enters the configuration network, so each draw gives the input its own atom positions, detunings and tethers and one ensemble member $\hat{y}_m$. Training minimizes the fair energy score (Gneiting & Raftery, 2007), $\mathrm{ES} = \frac{1}{M}\sum_m \|\hat{y}_m - y\| - \frac{1}{M(M-1)}\sum_{m<m'} \|\hat{y}_m - \hat{y}_{m'}\|$ (M=2, norms in VRMSE units), and a validation-tuned scale τ sets the spread (Appendix O).

4 Protocol

Hypothesis and kill rule. H1: the register improves the benchmark metric over a parameter-matched classical core trained with identical hyperparameters and step budget; if the core is statistically indistinguishable or better across seeds, H1 is rejected for that regime (rule fixed before each round). Each benchmark has one primary row fixed a priori. Tuning (Optuna, train/validation only) gives every model, the backbone included, the same trial budget; TEST is scored once per final configuration.

Statistics. Per row, seed-paired relative gains, paired t tests and the minimum detectable effect; across rows, a tie-aware sign test and a random-effects pool (Table 4); on TRL2D, per-trajectory Wilcoxon tests, Holm-corrected. $\pm$ is the standard deviation over three seeds.

Pre-registration. Hypotheses, kill rules and decision rules were timestamped before the corresponding runs (self-attested for review; code, logs and the record are released); the one extension registered after its models' results were known (§5.2) is labelled as such.

5 Results

5.1 The register ties a matched classical core across eight benchmarks

Table 1 reads left to right: published state of the art, our classical model, and the same model with a matched MLP or with qubits. The rows use the released checkpoints of The Well's CNextU-net (Ohana et al., 2024) and of TokaMind (Boschi et al., 2026) on TokaMark (Rousseau et al., 2026), the official U-Net pipelines of a groundwater heat-transport surrogate (Pelzer et al., 2025) and of the Fusion Equilibrium Challenge (Waller et al., 2026), a TORAX benchmark (Citrin et al., 2024), and our FNO arena. Classical + qubits beats the published state of the art on shear flow, convection, TRL2D and RealPDEBench and matches it on TokaMark (0.1404 vs 0.140); most of the margin comes from the classical part, and routed qubits add 12–13% on shear flow and 5–6% on convection, where the error is predictable (Figure 2). In the neck no comparison is significant after correction (sign test $p = 0.69$ against the matched core; Figure 7), because a pooled register sees only the spatial mean of the error (Proposition 1). Where a fine-tuned model beats a published number, the backbone alone beats it within seed noise, so those gains belong to fine-tuning (Appendix F); four neck rows are underpowered (minimum detectable effects 53–99%, Appendix A).

5.2 A candidate effect that did not replicate: long-horizon rollouts on TRL2D

On TRL2D we fine-tuned the official CNextU-net in two registered rounds. With a one-step objective every model improves on the checkpoint indistinguishably (backbone 0.19307 ± 0.00008, MLP 0.19314 ± 0.00019, register 0.19315 ± 0.00009). Trained through rollouts (Brandstetter et al., 2022) and scored on The Well's long-horizon window 13:30 (published 1.01), the register was the only model below the checkpoint on three seeds ($\mathbf{0.9528 \pm 0.0542}$ vs 1.0301 ± 0.0449 and MLP 1.0322 ± 0.0259), but a registered fourth seed did not replicate the edge (Table 10, Appendix H).

Table 1: From the published state of the art to classical + qubits (test split; lower is better except S; three seeds). Top: the routed co-processor on a fixed classical part (fine-tuned CNextU-net, frequency gate, classical head), one-step VRMSE; persistence scores 0.0646, 0.2024, 0.5761 and 0.418 on these rows. Bottom: the standard pooled register in the neck of each model (audit protocol; seeds on which the qubits are better in parentheses; †two length-matched seeds; ‡backbone tuned under an earlier objective). Full statistics: Table 4.

Benchmark	Published	Our classical	+MLP	+30 qubits	+50 qubits	vs MLP	vs classical
Routed co-processor: input-dependent qubit configuration							
shear flow	0.8080	0.0236	0.0213	0.0207	**0.0205**	+2.6%	+12.3%
convection (RB)	0.6566	0.0449	0.0428	0.0426	**0.0422**	+0.5%	+5.1%
turbulence (TRL2D)	0.1956	**0.1922**	**0.1922**	**0.1922**	—	0.0%	0.0%
MHD (MHD_64)	**0.1633**	0.1733	0.1733	0.1733	—	0.0%	0.0%
Standard design: pooled register in the neck							
turbulence (TRL2D arena)	—	0.9567	0.9534	**0.8447**	—	+11.4% (3/3)	+11.7% (2/3)
MHD (MHD_64)†	**0.1633**	0.1897	0.1897	0.1896	—	+0.0% (2/2)	+0.0% (2/2)
convection (RB)	0.6566	0.4563	0.5275	**0.4221**	—	+20.0% (3/3)	+7.5% (1/3)
shear flow	0.8080	0.2686	**0.2539**	0.2659	—	−4.7% (2/3)	+1.0% (2/3)
tokamak transport (TORAX)	—	0.0158	**0.0141**	0.0168	—	−19.2% (0/3)	−6.4% (2/3)
geothermal plume (1HP)‡	—	0.0291	0.0292	**0.0288**	—	+1.3% (2/3)	+1.2% (2/3)
fusion equilibrium ($S\uparrow$)‡	**0.9959**	0.7905	0.7848	0.7929	—	+1.0% (2/3)	+0.3% (1/3)
tokamak signals (TokaMark)	**0.140**	0.1509	0.1504	0.1501	—	+0.2% (1/3)	+0.5% (2/3)
Neck rows better / worse / tied (sign test)						4 / 2 / 2 (p=0.69)	5 / 1 / 2 (p=0.22)

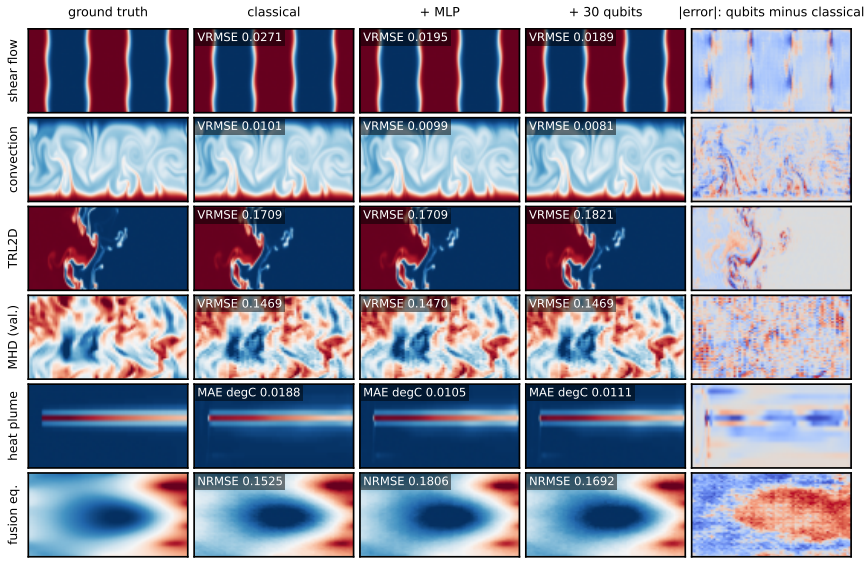

Figure 2: Model outputs against ground truth, one sample per benchmark fixed in advance (MHD: validation; heat plume: °C, classical = pretrained U-Net; fusion: flux $\psi(R, Z)$, NRMSE). Right: error change from classical to +30 qubits (blue: lower). Qubits cut the error on shear layers, convection plumes and the heat plume; MHD keeps the backbone's checkerboard, and on this fusion frame the classical model is best.

5.3 Why ties dominate: pooling caps utilization at the DC mode

Proposition 1 (a pooled register sees only the spatial mean)**.** *Let $P(z) = \frac{1}{N}\sum_{s=1}^{N} z_s$ be mean pooling over N positions and let g satisfy $\sum_s g_s = 0$. Then $P(z+g) = P(z)$; in particular P is invariant to adding any multiple of a non-DC discrete Fourier mode ($\zeta^N = 1$, $\zeta \neq 1$). A spatially constant write-back c is orthogonal to every zero-mean field, $\sum_s c\, g_s = 0$. Hence the insertion of*

Table 2: Correcting only the low-frequency or high-order part of a model's prediction (test split unless noted; three seeds). RealPDEBench airfoil: relative L_2 error on 5,140 test windows; the corrector reads and writes the lowest $3\times4\times4$ space–time Fourier modes of the model's prediction. Fusion equilibrium (Sophelio): official score S on validation; the corrector writes only PCA components $k \geq 8$ of a frozen model.

Model (benchmark)	Model alone	+linear	+MLP	+qubits
U-Net, released (RealPDEBench) ↓	0.00948	**0.00937**	0.00939	0.00939
FNO, released (RealPDEBench) ↓	0.01413	0.01292	**0.01267**	**0.01267**
frozen NAS model (Sophelio, validation) ↑	0.9864	—	0.9876	0.9879

§3 reads and writes only the DC spatial mode of the latent it is attached to, whatever its core. In any orthonormal basis, even a perfect correction of the coefficients in a set T leaves exactly the error energy outside T.

All four statements are proved in Lean 4 with Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020), standard axioms only (Appendix B).

Call the share of the error energy in the modes an insertion can read and write its *utilization ceiling*; for pooled insertions it is the DC share. On TRL2D (Table 5, Appendix B) the DC shell holds **2.9%** of the fine-tuned model's one-step error energy; the backbone misses 8–11% of the one-step change at $1 \leq |k| \leq 4$ but 34–45% at $|k| \geq 17$, the spectral bias of neural surrogates (Rahaman et al., 2019; Lippe et al., 2023). Paired over seeds, the register's only measurable effect is at $k{=}0$ ($-0.49\% \pm 0.21$ of that band's error, 3/3 seeds; matched MLP $+0.10\% \pm 0.19$); every other band moves within noise. The classical core is confined to the same mode, which is why the two tie.

5.4 Raising utilization: aiming the insertion where the error lives

If pooling is capped by the DC share, an insertion aimed where the error lives should help, and help any core (Table 2). On the RealPDEBench airfoil (Hu et al., 2026), a low-mode corrector lowers the released U-Net's test relative L_2 from 0.00948 to 0.00937 and the FNO's by 10%; a linear map captures most of it, and the register equals its matched MLP to four digits. On the Fusion Equilibrium Challenge (full-data pipeline), a coupling that writes only the fine PCA components lifts VALID S by $+0.0012$ to $+0.0015$ with either core; the register leads its matched MLP by $+0.0002$ ($p \geq 0.41$), or by $+0.0010$ ($p = 0.09$ corrected) where the learning rate was selected for the register arm (Appendix C).

Evaluation error, persistence, and a frequency-aware gate. On rayleigh_benard and shear_flow, persistence beats every published model and every backbone (Fig. 3a): the benchmark's per-window VRMSE is dominated by low-variance windows that MSE-trained backbones barely weight. A causal gate between backbone and persistence per Fourier octave (its optimum is proved never worse than either) cuts shear_flow's test VRMSE from 0.2297 to 0.0272. Routing then sends the reachable error to the qubits (Fig. 3b), and the register, a GNN on the same graph and an MLP all reach about 0.020 (Appendix G). Trained end to end with the metric as loss, a three-backbone gated ensemble reaches 0.0403 on rayleigh_benard (full test; persistence 0.2024).

5.5 Which input-dependent allocation? A registered cross-benchmark study

Design. To separate *where* the atoms work from *what* computes there, we freeze each backbone and attach a co-processor to its residual $r = y - p$ in an orthonormal multi-level Haar basis. Per input, a strategy picks the K tiles the atoms read and write: *pooled* (the audit's insertion), *static*, *proxy*, *learned*, *gain* (the budgeted optimum) or *random* (Appendix D). Atoms sit at their tiles, so the allocation sets the interaction graph, and every core starts as the same linear corrector. Proved bounds fix what can be measured: the captured share never exceeds the *utilization ceiling* U, ceilings nest, and a broadcast correction removes at most the DC energy. Rows: TRL2D, rayleigh_benard, shear_flow and MHD_64 with the released CNextU-net, RealPDEBench with its released U-Net and FNO, and the heat plume with the 1HP_NN U-Net; $K = 30$, three seeds, every choice on validation, test once; the protocol and its two amendments were registered before the runs.

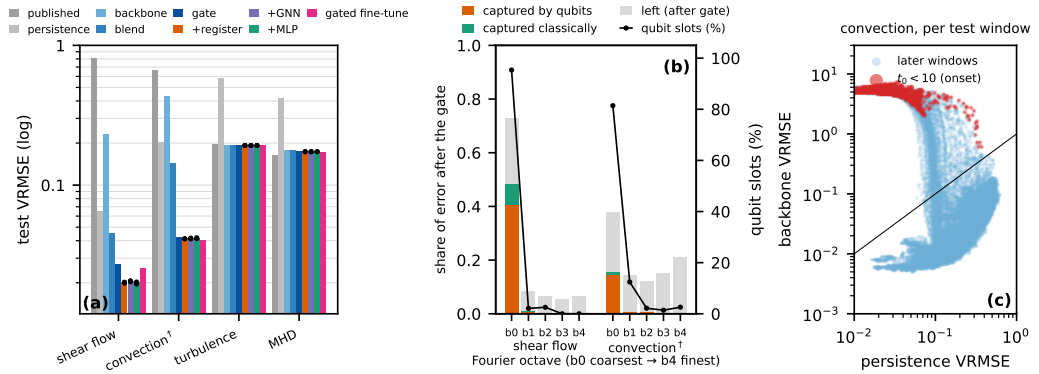

Figure 3: Where the error sits on the four Well benchmarks (shear flow, Rayleigh–Bénard convection, turbulent radiative layer, 3-D MHD). (a) Test VRMSE, log scale, from the published best model through persistence ($\hat{y}_{t+1}{=}x_t$), our backbone, the frequency gate, and the gate plus qubits, GNN or MLP (dots: seeds; †stratified test sample). (b) After the gate, the share of the remaining error in each Fourier octave captured by the qubits, by the classical head, or left; the line gives the router's share of the 30 qubits per octave: 80–95% go to the coarse octave that holds the predictable error. (c) Convection, per test window: at convection onset (red) the backbone is 10–100× worse than persistence, which the gate corrects.

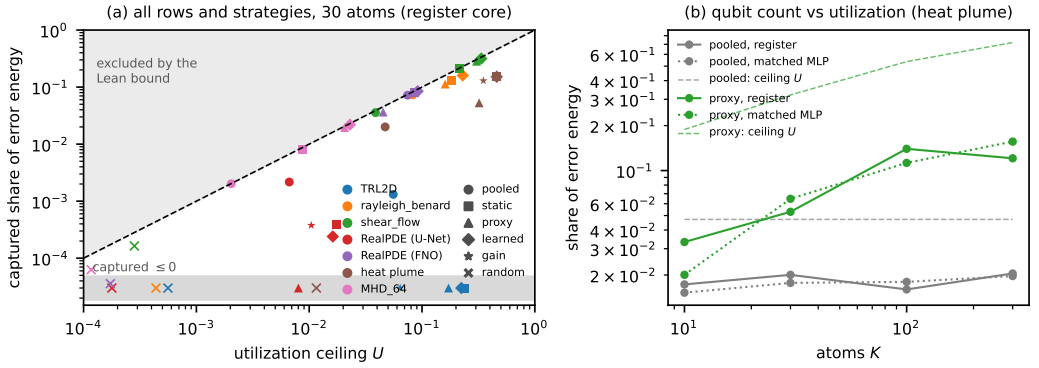

Figure 4: Allocation study (frozen backbones, test split, three seeds). (a) Captured share of each backbone's error energy against the utilization ceiling U of the allocation, for every row and strategy with the register core; no point can lie above the diagonal. (b) On the heat plume, adding atoms to the pooled insertion changes nothing (its ceiling is the DC share); routed atoms keep improving as their ceiling grows. Cores coincide throughout (dotted: matched MLP).

Allocation sets the ceiling; predictability sets how much of it is reached. With the same 30 atoms, routing raises the ceiling far above the pooled DC share (Table 3; e.g. from 0.039 to 0.34 on shear_flow and from 0.047 to 0.46 on the heat plume), and where the backbone's error is systematic the gain follows: routed atoms cut mean VRMSE by 36% on rayleigh_benard and 26% on shear_flow against about 7% pooled (both relative to released backbones that persistence beats in the mean, Appendix G); on the heat plume 30 static atoms lower error by 9.2% against 1.45% pooled (oracle 15%), and the matched MLP does as well; RealPDEBench's FNO reaches 93% of U. On TRL2D nothing is captured by any strategy or core, even under oracle placement with a 44% ceiling. Yet with the residual supplied as a feature the same pipeline captures 99.7% of it, so the released turbulence backbone's fine-scale error is not predictable by these correctors from causal inputs. The registered predictions split accordingly (captured never exceeded U; captured rises with U across the 42 row–strategy cells, Spearman $\rho = 0.71$ with row-bootstrap 95% CI $[0.18, 0.98]$; but routing beats pooling on 5 of 7 rows, failing exactly on the two with unpredictable error, and routing by the

Table 3: Allocation study, 30 atoms, register core, test split, mean of three seeds. Pooled: the audit's insertion (ceiling = DC share). Best: the register's validation-best tile strategy, its utilization ceiling U, the captured share of error energy and the test-metric gain over the frozen backbone (VRMSE; relative L_2 on RealPDEBench); the last column is the register's seed-paired gain over its matched MLP at that strategy. On TRL2D and RealPDEBench (U-Net) every tile strategy captures at most 0.0004, and the listed one is the tie-broken validation choice. rayleigh_benard's mean gain concentrates in the convection-onset transient (median VRMSE -4%). †Released checkpoint does not reproduce its published number under our evaluator (Appendix F).

	pooled		best tile strategy				
Row (frozen backbone)	DC share	gain	strategy	U	captured	gain	reg. vs MLP
TRL2D (CNextU-net)	0.056	$+0.06\%$	static	0.238	0.000	$+0.00\%$	$+0.00\%$ (0/3)
rayleigh_benard (CNextU-net)	0.083	$+6.69\%$	proxy	0.161	0.113	$+36.25\%$	-0.48% (1/3)
shear_flow (CNextU-net)†	0.039	$+6.77\%$	learned	0.335	0.316	$+26.29\%$	$+0.03\%$ (3/3)
MHD_64 (CNextU-net)†	0.002	$+0.11\%$	gain	0.023	0.022	$+0.87\%$	$+0.00\%$ (1/3)
RealPDEBench (U-Net)	0.007	$+0.12\%$	static	0.017	0.000	$+0.01\%$	-0.00% (0/3)
RealPDEBench (FNO)	0.074	$+3.32\%$	gain	0.091	0.085	$+3.87\%$	$+0.01\%$ (3/3)
Heat plume (1HP_NN U-Net)	0.047	$+1.45\%$	static	0.459	0.152	$+9.23\%$	-0.18% (1/3)

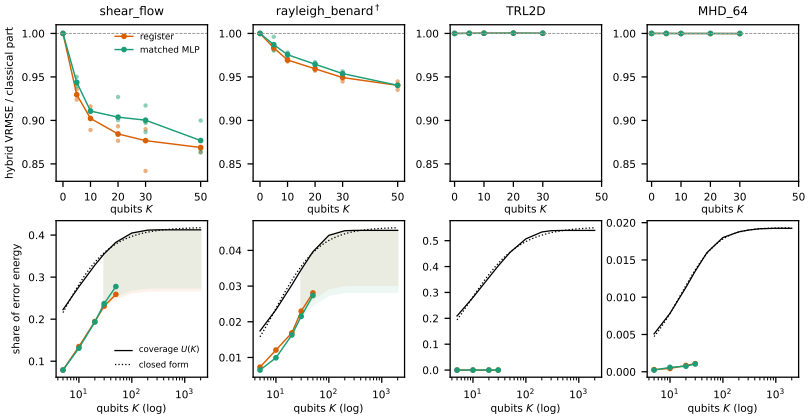

Figure 5: Hybrid scaling with the qubit budget, classical part fixed (three seeds; dots are seeds). Top: test VRMSE relative to the classical part ($K{=}0$). Error falls with each added atom group on shear_flow and rayleigh_benard, where the backbone's error is predictable, and stays at 1 on TRL2D and MHD_64, where it is not. Bottom: captured share (register, MLP) against the reachable share $U(K)$ (solid) and the closed form of $G(K)$ (dotted); the 50-qubit points were predicted before they ran.

co-processor's own predicted gain beats energy routing on only 3 of 7, being within noise of learned and proxy routing).

Qubit count versus utilization. Figure 4b shows this on the heat plume: 10 to 300 pooled atoms capture 1.5–2.1% (ceiling 4.7%), while routed atoms capture 2, 6, 11 and 16% as their ceiling grows from 19% to 72%. With the classical part fixed and positions, detunings, routing and tether strengths set per input, the hybrid's error falls steadily with the budget, by 13% on shear_flow and 6% on rayleigh_benard at 50 qubits (12% and 5% at 30; the matched MLP 10% and 5%). This is the hybrid scaling law $G(K) = \eta U_\infty[1-(1+K/K_0)^{1-a}]$: the reachable share $U(K)$ is computed from test data without fitting and follows the closed form within 0.015 (a 1.85–2.62); the registered 50-qubit prediction held on all six seeds (Figure 5, Appendix N). A registered control shows the routing carries the gain: with the same routing, a fixed layout without tethers matches per-input positions and tethers within 0.1%, and a 3-D layout does not help (Appendix N).

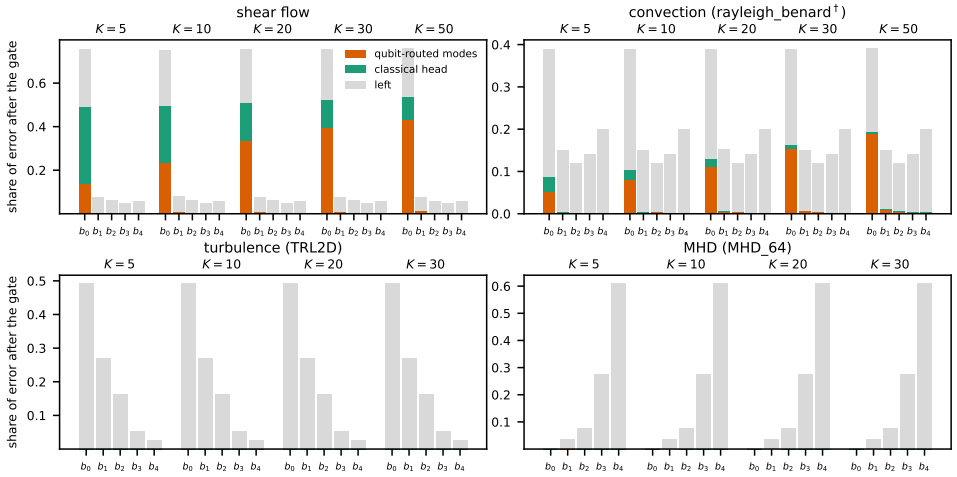

Figure 6: Where added qubits work, per Fourier octave (b_0 coarsest to b_4 finest), register arm, three seeds. Bars: error after the gate captured by the qubit-routed modes (orange), by the classical head (green), and left (grey). On shear flow and convection the error sits in the coarse octave and the qubits take it over from the classical head as K grows; on TRL2D it is spread across octaves and on MHD it sits in the finest octave, and no budget captures it.

The core. At each row's validation-best strategy the register ties its matched MLP (3 rows better, 3 worse, 1 tied; random-effects $+0.006\%$ $[-0.005, +0.017]$), as it does the mean-field register and a $4\times$ MLP; a register placed beside an MLP on the same atoms ties two MLPs (-0.43% to $+0.01\%$ on the five rows with predictable error, none significant). A controlled probe finds no analog edge, and on targets generated by interacting Rydberg dynamics the register beats matched MLPs and GNNs but not its classical mean-field twin (Appendix D). With configurations drawn per input from a noise code and trained on the fair energy score, every hybrid's energy score falls 16–20% below its deterministic version, while the ensemble mean's VRMSE does not improve; the register beat its matched MLP on shear_flow (6/6 seeds, Holm $p = 0.044$) but not an MLP one hidden unit wider (4/6, $p = 0.24$; Appendix O).

6 Limitations and Conclusion

Three seeds bound every claim, with four audit rows underpowered for $\leq 20\%$ effects. Registers are simulated exactly in groups of at most ten atoms (Liu et al., 2025); decoherence is unmodelled (readout and calibration tests: Appendix M). The allocation study freezes backbones and evaluates test samples on four rows.

Within these limits, the evidence supports one conclusion. A hybrid model should budget utilization, not qubits: atoms aimed at the error a backbone gets wrong predictably turn a fixed budget into gains that grow with budget, whereas pooled atoms cannot, whatever their number; we prove this and measure it on seven rows. At sizes that can be simulated exactly, a matched classical core does that job as well as the register; the one candidate regime, long-horizon rollouts under rollout-aware training, did not replicate at four seeds. Next are long-horizon world models for industrial digital twins (plasma control, geothermal reservoirs, fleet routing, video; Li et al., 2025, Appendix Q), where per-step error compounds and the law sizes a qubit budget in advance, and runs past exact simulation on hundreds of atoms in Pasqal- and QuEra-class processors, testing whether native Hamiltonian dynamics raise η where classical cores cannot follow. None of these results is a claim of quantum advantage (Schuld & Killoran, 2022; Cerezo et al., 2025).

AI USE STATEMENT

In this work we used a generative AI coding agent (Anthropic's Claude Code) for code implementation, running and monitoring experiments, computing statistics, producing figures, literature search, and drafting and revising the manuscript text, under the authors' direction and written protocols; and a second agent (Google's Gemini, via the Antigravity CLI) for an independent leakage review of code, a reviewer-style critique of the manuscript, and one validation-only post-processing sweep, whose outputs the first agent checked before use. We did not use generative AI tools to set hypotheses, pre-registered decision rules, kill rules, or benchmark selection, which were fixed by the authors. We have reviewed all AI-assisted work: every reported number regenerates from released scripts and result files and was cross-checked by an independent audit script (`paper/judge/judge.py`); every citation was verified against the publisher's or arXiv record; generated code is covered by unit tests. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The work uses public simulation and measurement datasets under their licenses (RealPDEBench under CC BY-NC, used for non-commercial research) and a benchmark we generated with an open-source simulator; no human subjects or personal data are involved. The application domains (fusion, geothermal and flow modelling) are civilian energy research. All experiments ran on one workstation; negative results are reported in full to discourage selective reporting of quantum benefits.

REPRODUCIBILITY STATEMENT

Section 4 and Appendices A–J specify every protocol, split, metric and statistic. The code release contains the register implementation with unit tests, checkpoint-surgery and fine-tuning code, evaluation scripts for every benchmark, the audit script that regenerates Table 1 from per-seed result files, the Lean 4 sources of Proposition 1, the timestamped pre-registration record, training logs and evaluation arrays, and the package lock file; released checkpoints and datasets are public. The anonymized supplementary material already contains the code, the Lean sources, every per-seed result file behind the allocation, scaling, probabilistic and routing tables, the pre-registration notes, and a script that regenerates those tables.

REFERENCES

Vytautas Abramavicius, Melvin Mathé, Gergana V. Velikova, João P. Moutinho, Mario Dagrada, Vincent E. Elfving, Alexandre Dauphin, Joseph Vovrosh, and Roland Guichard. PulserDiff: A pulse differentiable extension for Pulser. *arXiv preprint arXiv:2505.16744*, 2025.

Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, et al. Skillful joint probabilistic weather forecasting from marginals. *arXiv preprint arXiv:2506.10772*, 2025.

Daniel Barredo, Vincent Lienhard, Sylvain de Léséleuc, Thierry Lahaye, and Antoine Browaeys. Synthetic three-dimensional atomic structures assembled atom by atom. *Nature*, 561:79–82, 2018.

Federico Berto, Chuanbo Hua, Nayeli Gast Zepeda, André Hottung, Niels Wouda, Leon Lan, et al. RouteFinder: Towards foundation models for vehicle routing problems. *Transactions on Machine Learning Research*, 2025.

Black Forest Labs. FLUX 3 Action: open-weights world action model (model card and documentation). `https://huggingface.co/black-forest-labs/flux-3-action-droid`, `https://docs.bfl.ai/flux_3/flux3_action_overview`, 2026.

Tobia Boschi, Andrea Loreti, Nicola C. Amorisco, Rodrigo H. Ordonez-Hurtado, Cécile Rousseau, George K. Holt, Eszter Székely, Alexander Whittle, Samuel Jackson, Adriano Agnello, Stanislas Pamela, Alessandra Pascale, Robert Akers, Juan Bernabe Moreno, Vassil Alexandrov, and Mykhaylo Zayats. TokaMind: A multi-modal transformer foundation model for tokamak plasma dynamics. *arXiv preprint arXiv:2602.15084*, 2026.

Joseph Bowles, Shahnawaz Ahmed, and Maria Schuld. Better than classical? The subtle art of benchmarking quantum machine learning models. *arXiv preprint arXiv:2403.07059*, 2024.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations (ICLR)*, 2022.

Rodrigo Araiza Bravo, Khadijeh Najafi, Xun Gao, and Susanne F. Yelin. Quantum reservoir computing using arrays of Rydberg atoms. *PRX Quantum*, 3:030325, 2022.

Ferhat Ozgur Catak. Measurement-budget allocation in quantum learning with finite-shot generalization guarantees. *arXiv preprint arXiv:2608.24891*, 2026.

M. Cerezo, Martin Larocca, Diego García-Martín, N. L. Diaz, Paolo Braccia, Enrico Fontana, Manuel S. Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut Thanasilp, Eric R. Anschuetz, and Zoë Holmes. Does provable absence of barren plateaus imply classical simulability? *Nature Communications*, 16:7907, 2025.

Hao-Yuan Chen. Investigating quantum-embedded transformers on classical datasets for cross-modality classification. *arXiv preprint arXiv:2608.06846*, 2026.

Jonathan Citrin, Ian Goodfellow, Akhil Raju, Jeremy Chen, Jonas Degrave, Craig Donner, Federico Felici, Philippe Hamel, Andrea Huber, Dmitry Nikulin, David Pfau, Brendan Tracey, Martin Riedmiller, and Pushmeet Kohli. TORAX: A fast and differentiable tokamak transport simulator in JAX. *arXiv preprint arXiv:2406.06718*, 2024.

Robert de Keijzer, Luke Visser, Oliver Tse, and Servaas Kokkelmans. Consensus-based qubit configuration optimization for variational algorithms on neutral atom quantum systems. *npj Quantum Information*, 2025.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, volume 12699 of *Lecture Notes in Computer Science*, pp. 625–635, 2021.

Nicholas S. DiBrita, Daniel Leeds, Yuqian Huo, Jason Ludmir, and Tirthak Patel. ReCon: Reconfiguring analog rydberg atom quantum computers for quantum generative adversarial networks. In *Proceedings of the International Conference on Computer-Aided Design (ICCAD)*, 2024.

Nicholas S. DiBrita, Jason Han, and Tirthak Patel. ResQ: A novel framework to implement residual neural networks on analog rydberg atom quantum computers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Elies Gil-Fuster, Casper Gyurik, Adrián Pérez-Salinas, and Vedran Dunjko. On the relation between trainability and dequantization of variational quantum learning models. In *International Conference on Learning Representations (ICLR)*, 2025.

Syeda Anshrah Gillani, Mirza Samad Ahmed Baig, Shahid Munir Shah, Asher Ali, and Hamzah Siddiqui. How quantum is the advantage? a fair, calibration- and noise-aware benchmark and attribution audit of quantum machine learning for network intrusion detection. *arXiv preprint arXiv:2608.18155*, 2026.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Felix Herbort, Ellen Sarauer, Daniel Ohl de Mello, Paul Christiansen, Steffen Hien, Cedric Brügmann, Dieter Jaksch, Veronika Eyring, Martin Kiffner, and Mierk Schwabe. How hard is quantum advantage? a cloud microphysics stress test for variational quantum models. *arXiv preprint arXiv:2607.04915*, 2026.

Md Aminur Hossain, Ayush V. Patel, Siddhant Gole, Sanjay K. Singh, and Biplab Banerjee. HQF-Net: A hybrid quantum-classical multi-scale fusion network for remote sensing image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2026.

Peiyan Hu, Haodong Feng, Hongyuan Liu, Tongtong Yan, Wenhao Deng, Tianrun Gao, Rong Zheng, Haoren Zheng, Chenglei Yu, Chuanrui Wang, Kaiwen Li, Zhi-Ming Ma, Dezhi Zhou, Xingcai Lu, Dixia Fan, and Tailin Wu. RealPDEBench: A benchmark for complex physical systems with real-world data. *arXiv preprint arXiv:2601.01829*, 2026.

Naman Jain and Amir Kalev. QuFeX: Quantum feature extraction module for hybrid quantum-classical deep neural networks. *Quantum Science and Technology*, 11:015017, 2026.

Hailong Jiang, Emran Hossain, Feng Yu, Jianfeng Zhu, Guilin Zhang, and Wulan Guo. Quantum-structured world models (QSWMs) for predictive latent dynamics. *arXiv preprint arXiv:2608.05371*, 2026.

Haiyue Kang, Martin Sevior, and Muhammad Usman. Fourier analysis of quantum neural network with non-linear data embedding. *arXiv preprint arXiv:2606.14206*, 2026.

Hedwig Keller, Mierk Schwabe, and Veronika Eyring. Shadow models of a quantum model for cloud cover and the influence of finite sampling noise. *arXiv preprint arXiv:2608.20076*, 2026.

Sang Hyub Kim, Jonathan Mei, Claudio Girotto, Masako Yamada, and Martin Roetteler. Quantum large language model fine-tuning. *arXiv preprint arXiv:2504.08732*, 2025.

Toshiaki Koike-Akino, Francesco Tonin, Yongtao Wu, Frank Zhengqing Wu, Leyla Naz Candogan, and Volkan Cevher. Quantum-peft: Ultra parameter-efficient fine-tuning. In *International Conference on Learning Representations (ICLR)*, 2025.

Milan Kornjača, Hong-Ye Hu, Chen Zhao, Jonathan Wurtz, Phillip Weinberg, Majd Hamdan, et al. Large-scale quantum reservoir learning with an analog quantum computer. *arXiv preprint arXiv:2407.02553*, 2024.

Pranav Lakshmanan and Paras Chopra. Hybrid neural world models. *arXiv preprint arXiv:2605.28317*, 2026.

Diana Legziel Levy, Menachem Finkelstein, Peter Chin, Eilon Vaadia, and Sarel Cohen. Quantum-inspired hybrid neural networks for neural decoding: A controlled ablation study of learnable quantum sidecar integration. *arXiv preprint arXiv:2608.22475*, 2026.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Chen-Yu Liu, Chao-Han Huck Yang, Hsi-Sheng Goan, and Min-Hsiu Hsieh. A quantum circuit-based compression perspective for parameter-efficient learning. In *International Conference on Learning Representations (ICLR)*, 2025.

Hannah J. Manetsch, Gyohei Nomura, Elie Bataille, Kon H. Leung, Xudong Lv, and Manuel Endres. A tweezer array with 6100 highly coherent atomic qubits. *Nature*, 647:60–67, 2025.

Paolo Marcandelli, Yuanchun He, Stefano Mariani, Martina Siena, and Stefano Markidis. Partitioned hybrid quantum fourier neural operators for scientific quantum machine learning. *arXiv preprint arXiv:2507.08746*, 2025.

Michael McCabe, Payel Mukhopadhyay, Tanya Marwah, Bruno Regaldo-Saint Blancard, Francois Rozet, Cristiana Diaconu, Lucas Meyer, Kaze W. K. Wong, Hadi Sotoudeh, Alberto Bietti, Irina Espejo, Rio Fear, Siavash Golkar, Tom Hehir, Keiya Hirashima, Geraud Krawezik, Francois Lanusse, Rudy Morel, Ruben Ohana, Liam Parker, Mariel Pettee, Jeff Shen, Kyunghyun Cho, Miles Cranmer, and Shirley Ho. Walrus: A cross-domain foundation model for continuum dynamics. *arXiv preprint arXiv:2511.15684*, 2025.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Keaton Burns, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. The well: A large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 2024.

Tushar Pandey. When classical baselines are tuned as carefully as the quantum model, does quantum reservoir computing still win? *arXiv preprint arXiv:2607.09905*, 2026a.

Tushar Pandey. A quantum reservoir architecture for chaotic forecasting and a test of whether its high dimension helps. *arXiv preprint arXiv:2607.07978*, 2026b.

Mateusz Papierz, Asel Sagingalieva, Alix Benoit, Toni Ivas, Elia Iseli, and Alexey Melnikov. Hybrid fourier neural operator for surrogate modeling of laser processing with a quantum-circuit mixer. *arXiv preprint arXiv:2604.04828*, 2026.

Julia Pelzer, Corné Verburg, Alexander Heinlein, and Miriam Schulte. Resolving extreme data scarcity by explicit physics integration: An application to groundwater heat transport. *arXiv preprint arXiv:2507.06062*, 2025.

Pierre Pinson and Julija Tastu. Discrimination ability of the Energy score. Technical Report Technical Report 2013-15, Technical University of Denmark, DTU Compute, 2013.

Samuele Punzo, Niccolò Caselli, Ippokratis Pantelidis, Francesco Massafra, Salvatore Lo Sardo, and Mohammadreza Salehi. How do video foundation models encode intuitive physics? probing across pretraining paradigms. *arXiv preprint arXiv:2606.09646*, 2026.

Jun Qi, Chao-Han Huck Yang, Pin-Yu Chen, Min-Hsiu Hsieh, Hector Zenil, and Jesper Tegner. Quantum LEGO learning: A modular design principle for hybrid artificial intelligence. *arXiv preprint arXiv:2601.21780*, 2026.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Snehal Raj and Brian Coyle. Quic: Quantum-inspired compound adapters for parameter efficient fine-tuning. *arXiv preprint arXiv:2502.06916*, 2025.

Stephan Rasp, Boris Babenko, Dominic Masters, Andrew El-Kadi, Samier Merchant, et al. WeatherNext 3: Increasing resolution and performance of global weather models with raw observations. *arXiv preprint arXiv:2609.03582*, 2026.

Cécile Rousseau, Samuel Jackson, Rodrigo H. Ordonez-Hurtado, Nicola C. Amorisco, Tobia Boschi, George K. Holt, Andrea Loreti, Eszter Székely, Alexander Whittle, Adriano Agnello, Stanislas Pamela, Alessandra Pascale, Robert Akers, Juan Bernabe Moreno, Sue Thorne, and Mykhaylo Zayats. TokaMark: A comprehensive benchmark for MAST tokamak plasma models. *arXiv preprint arXiv:2602.10132*, 2026.

Shyam Sankaran, Hanwen Wang, and Paris Perdikaris. Small models, strong priors: Architectural inductive bias for parameter-efficient neural pde solvers. *arXiv preprint arXiv:2605.25949*, 2026.

Franz J. Schreiber, Jens Eisert, and Johannes Jakob Meyer. Classical surrogates for quantum learning models. *Physical Review Letters*, 131:100803, 2023.

Maria Schuld and Nathan Killoran. Is quantum advantage the right goal for quantum machine learning? *PRX Quantum*, 3:030101, 2022. doi: 10.1103/PRXQuantum.3.030101.

Jiwoo Son, Zhikai Zhao, Federico Berto, Chuanbo Hua, Zhiguang Cao, et al. RRNCO: Towards real-world routing with neural combinatorial optimization. In *International Conference on Learning Representations*, 2026.

Melvin Strobl, Maja Franz, Lukas Scheller, Eileen Kuehn, Wolfgang Mauerer, and Achim Streit. Beyond gates: Pulse level quantum Fourier models. *arXiv preprint arXiv:2605.04945*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP)*, pp. 367–381, 2020.

Danil Vyskubov, Kirill Vyskubov, Nouhaila Innan, and Muhammad Shafique. Scaling laws for hybrid quantum neural networks: Depth, width, and quantum-centric diagnostics. *arXiv preprint arXiv:2604.06007*, 2026.

Matthew Waller, Craig Michoski, Tapan Ganatma Nakkina, Brian Sammuli, William Boyes, Mitchell Clark, Sterling Smith, and Raffi Nazikian. The fusion equilibrium challenge. Challenge starter kit and dataset, `https://github.com/Sophelio/fusion-equilibrium-challenge-starter`, 2026.

Maida Wang, Xiao Xue, Mingyang Gao, and Peter V. Coveney. Quantum-informed machine learning for predicting spatiotemporal chaos with practical quantum advantage. *Science Advances*, 12 (16), 2026a.

Peiyong Wang, Udaya Parampalli, and Casey R. Myers. Quantum spectral model: Data reuploading with input-conditioned frequency support. *arXiv preprint arXiv:2607.22516*, 2026b.

Ruocheng Wang, Zhuo Xia, Ge Yan, and Junchi Yan. QuanONet: Quantum neural operator with application to differential equation. In *International Conference on Machine Learning (ICML)*, 2025.

Xiao Xue, Maida Wang, Mingyang Gao, Minh Chung, and Peter V. Coveney. Explainable quantum-compressed machine learning for complex fluid flows. *arXiv preprint arXiv:2607.21688*, 2026.

Kaveh Zabihi, Hamid Montazeri, and Akke S. J. Suiker. Hybrid quantum-classical physics-informed neural networks for solving nonlinear PDEs: When and where hybridization is effective? *arXiv preprint arXiv:2606.04679*, 2026.

A AUDIT DETAILS

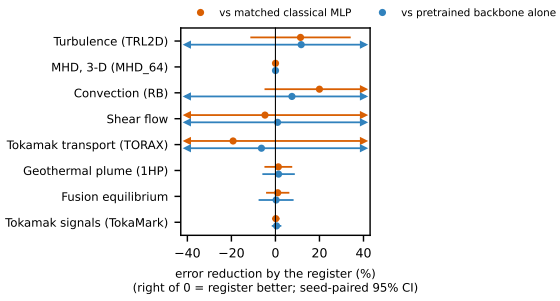

Figure 7: The audit at a glance: every 95% interval of the register's seed-paired gain crosses zero.

The audit script reads every per-seed value from the result files rather than from summaries, fixes one primary row per benchmark in advance, and reports per row the seed-paired gains, paired t, sign counts, d_z, Glass's Δ and the minimum detectable effect; across rows it reports a tie-aware sign test, the mean over rows with a t interval, and a DerSimonian–Laird random-effects pool with I^2 and leave-one-out values. Secondary rows: MHD_64 with all three seeds at unequal training length

Table 4: Matched-control audit, primary rows (three seeds unless noted; lower is better except Sophelio S). Gains are seed-paired relative improvements of the register (positive = better), with the number of seeds on which the register is better and the paired t-test p. Metrics: VRMSE (@8: eighth rollout step), MAE in °C, official score S, NRMSE. MDE: minimum detectable register-vs-backbone effect. Backbones: FNO (TRL2D arena), released CNextU-net (MHD_64, rayleigh_benard, shear_flow), 1-D spectral (TORAX-Bench), U-Net (heat plume, Sophelio starter), released TokaMind (TokaMark). †Two length-matched seeds (Appendix J). ‡Backbone tuned under the earlier single-seed objective (mixed protocol). On rayleigh_benard and shear_flow, persistence ($\hat{y}_{t+1} = x_t$) scores 0.2024 and 0.0646, below every backbone and adapter in this table; the fix is in Appendix G.

Benchmark	Backbone	+MLP	+Reg.	Reg. vs MLP (%)		Reg. vs backbone (%)	MDE (%)
TRL2D arena	0.9567	0.9534	0.8447	+11.4	3/3, .16	+11.7 2/3	53
MHD_64†	0.1897	0.1897	0.1896	+0.0	2/2, .28	+0.0 2/2	0.3
rayleigh_benard	0.4563	0.5275	0.4221	+20.0	3/3, .07	+7.5 1/3	69
shear_flow	0.2686	0.2539	0.2659	−4.7	2/3, .85	+1.0 2/3	99
TORAX-Bench	0.0158	0.0141	0.0168	−19.2	0/3, .40	−6.4 2/3	84
1HP heat plume‡	0.0291	0.0292	0.0288	+1.3	2/3, .44	+1.2 2/3	5.2
Sophelio $S\uparrow$‡	0.7905	0.7848	0.7929	+1.0	2/3, .46	+0.3 1/3	7.5
TokaMark 1-1	0.1509	0.1504	0.1501	+0.2	1/3, .63	+0.5 2/3	1.8
Rows better / worse / tied (sign-test p)				4 / 2 / 2 (p=0.69)		5 / 1 / 2 (p=0.22)	
Random-effects pooled gain [95% CI]				+0.53% [−0.39, +1.45]		+0.04% [−0.00, +0.08]	

(register vs MLP $+4.2\%$, 3/3, $p = 0.42$; flagged because four of nine runs stopped at about 7.5k of 20k steps and at matched length the arms agree to four decimals with the gate never opening, $|\alpha| \approx 4\times10^{-5}$); TORAX-Bench at fixed untuned hyperparameters (register vs backbone $+9.1\%$, 3/3, $p = 0.14$; vs MLP $+1.1\%$, 1/3); the heat plume at fixed hyperparameters (register vs MLP $+0.7\%$, 1/3); Sophelio at fixed hyperparameters (register vs MLP -0.2%, 1/3); TokaMark with a CNN backbone (one seed; register vs backbone $+4.0\%$); and the TRL2D arena with a shared-link register ($+13.3\%$ vs MLP, 3/3, $p = 0.20$). Three errors were caught by the script before reaching a claim: the unequal-length MHD comparison above; a shear-flow published anchor our evaluator does not reproduce (released checkpoint 0.9946 vs published 0.8080); and an inverse-variance pool that rows with identical arms dominate, which is why the mean over rows and leave-one-out values are reported. Register vs MLP leave-one-out pools range from $+0.16\%$ to $+2.16\%$, every interval including zero.

B SPECTRAL EVIDENCE AND MACHINE-CHECKED STATEMENTS

Table 5: TRL2D (180 test windows): where the released checkpoint's error lives, by FFT shell $|k|$, and what the register changes (paired over three Round-1 seeds, relative change of that shell's error energy vs vanilla).

$\|k\|$ shell	error energy share	change energy share	change missed by backbone	register vs vanilla
0 (DC)	0.029	0.022	0.23	$-0.49\% \pm 0.21$ (3/3)
1–4	0.201	0.361	0.08–0.11	within noise
5–16	0.428	0.459	0.11–0.21	within noise
17–32	0.231	0.116	0.34	within noise
≥33	0.111	0.042	0.44–0.45	within noise

Both decompositions conserve energy (Haar relative error 0, FFT shells 2×10^{-7}). With Haar wavelets the backbone misses 11% of the change at the coarsest scale, rising to 18, 26, 32 and 37% at 16, 8, 4 and 2 pixels; the top 30 of 196,416 detail tiles per sample hold 32.9% of the detail error energy, eleven times the DC ceiling, and a zero-parameter router that ranks tiles by the backbone's own predicted change captures 9.4% at K=30 (about 1,650 times a random choice); learned routers beat it only at $K \geq 3{,}000$. A routed register on TRL2D is a pilot (Section 6).

Machine-checked statements. Lean 4 with Mathlib, standard axioms only (`propext`, `Classical.choice`, `Quot.sound`), no `sorry`: `pool_blind` and `pool_blind_mode` (mean pooling is invariant to zero-mean perturbations, in particular non-DC Fourier modes); `broadcast_orthogonal`; `band_energy_split` and `residual_after_band_removal` (Parseval split and the exact residual after a perfect correction on a coefficient set, the ceiling of any K-tile corrector); `pointwise_min_optimal` and `swap_improves` (top-K by energy is the optimal budgeted assignment); `reupload_degree` (an L-layer re-uploading circuit's spectrum lies in $[-L, L]$); `vdw_sensitivity_anti` (geometry acts as frequency modulation); `lattice_sep`; and `dip_inplane`, `dip_axial`, `dip_magic` (signed dipolar couplings require an out-of-plane coordinate); and, for the allocation study (§5.5), `gain_le_ceiling` (in any orthonormal basis, a correction supported on a coefficient set T, with arbitrary coefficients, lowers the error energy by at most the error energy on T), `ceiling_mono` (nested allocations have nested ceilings), and `pooled_gain_le_dc` (a spatially constant correction c lowers $\sum_i e_i^2$ by at most $(\sum_i e_i)^2/n$, whatever core and however many atoms produced c); and, for the frequency-localized router and the persistence gate (Appendix F), `band_span_le_energy` (Bessel inside a band: the energy on orthonormal band-localized modes never exceeds the band's energy, so $\text{captured}_b \le \text{span}_b \le \text{full}_b$ per band and window), `gate_energy_exact` and `gate_coeff_exact` (the gated residual's energy and every mode coefficient are exact functions of three band statistics and two coefficients, for any gate value), and `gate_opt_le` with its specializations `gate_opt_le_persistence` and `gate_opt_le_trunk` (the best scalar gate per band is no worse than persistence or the backbone). The standard step that turns a parameterized circuit into a truncated Fourier series is not formalized. A separate file checks algebraic consequences of a proposed configuration-aware error model ($E = E_\infty + AP^{-\alpha}(1 - g\,u) + \dots$ with utilization $u \in [0, 1)$): improvement with utilization holds only when the signed gain $g \ge 0$, which is an empirical question the audit answers with a null.

C MOVING THE INSERTION: PROTOCOLS

RealPDEBench. Airfoil (NACA4418) real PIV velocity, 20 input and 20 output steps at 64×128; the benchmark's public splits (5,133 validation and 5,140 test windows over 98 trajectories), normalizer and metric code, with the full-set metric computed in exact chunks. The corrector reads the lowest $3 \times 4 \times 4$ space–time modes of the backbone's prediction and of the last input window, standardizes them on the training split, passes them through a zero-initialized gated core (none, linear, matched MLP, or $3{\times}10$-atom register; the MLP and register share input and output maps), and writes a correction to the same modes; it trains on training windows, early-stops on validation relative L_2, and scores test once per arm and seed. The corrector does not exceed the released paper's reported numbers, which are not reproducible on the current data release; the comparison is to the released checkpoint evaluated with the benchmark's code.

Fusion Equilibrium Challenge. Full-data pipeline (7,041 DIII-D shots, 1.56M frames; magnetics at 14 time lags and Thomson profiles), a PCA-64 flux decoder whose ground-truth projection scores $S = 0.9991$, and a stage-1 architecture search that fixed the frozen backbone. The stage-2 search over couplings (latent or spectral-routed, register size and graph, $\le$30 atoms) ran on VALID only; its top three configurations were re-run with three seeds beside matched MLP twins. Register vs MLP by configuration: 10-atom ring, learning rate $2.4{\times}10^{-3}$: 0.9878 vs 0.9876 ($p = 0.43$); same register at $1.6{\times}10^{-3}$: 0.9878 vs 0.9868 ($p = 0.030$, Bonferroni 0.09); 5-atom ring: 0.9879 vs 0.9876 ($p = 0.41$). After the single registered test scoring (0.9910 on our held-out split of training shots; the development leaderboard, which scores the public test fold, stood at 0.9932 in August and 0.9959 on September 25), a post-processing sweep on VALID only (shrinkage of high-order PCA coefficients, per-shot temporal smoothing, outlier filters) raised the full-VALID score of the entry ensemble from 0.99127 to 0.99143. That is 1.7 times the estimated ensemble seed SD, below the threshold we set in advance (3 times) for spending a second test scoring, so no second test was run.

Register (fixed backbone)	Atoms	VRMSE@8
FNO backbone alone	0	0.9567 ± 0.2554
+ Rydberg register	5	0.8427 ± 0.0709
+ Rydberg register	10	0.9332 ± 0.2166
+ Rydberg registers	20	0.9269 ± 0.2792
+ Rydberg registers (3×10)	30	0.8447 ± 0.0511
+ matched MLP (control)	—	0.9534 ± 0.1354
+ 30 atoms, conditioned geometry	30	0.8951 ± 0.1367
+ 30 atoms, mean-field pilot	30	0.8393 ± 0.0463

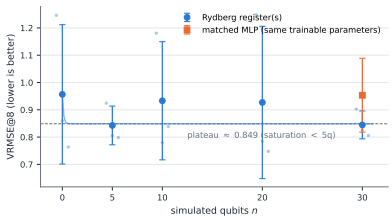

Figure 8: Rollout accuracy versus register size at fixed backbone, data and step budget (sprint arena, 3 seeds; amber: matched MLP). The 0- and 30-atom arms ran first; the interior arms and the mean-field pilot were pre-registered before execution. Points above 10 atoms add registers of 10. A constant fits the $n \geq 5$ points (a saturating fit puts any rise below five atoms).

D Allocation study: protocol, per-strategy results and the analog–digital probe

Qubit count and inertness. In a fixed-budget arena (FNO backbone, 2,000 steps), accuracy is flat from five qubits: 5 atoms reach 0.8427 ± 0.0709 against 0.8447 ± 0.0511 for 30 (Figure 8). The trained register is inference-inert (product-state surrogates reproduce it within $2.4{\times}10^{-5}$), so any gain arises in training (mechanism tests: Appendix H).

Caches and exactness. Each frozen backbone predicts every window once (fp32; the convolutional and spectral backbones have no layers NVFP4 applies to). Training windows are a uniform random sample of 3,000 (700 on the heat plume); validation and test windows are uniform samples of all windows or the full split. For every window we store the per-channel residual energy, the metric denominators, and, for 320 tiles per rule (proxy, static, random, oracle), the Haar coefficients of $p - x_t$, p, x_t, $x_t - x_{t-1}$ and r. Because the transform is orthonormal, the corrected per-channel error is $E_c - \sum_{i \in S} \left(r_{ic}^2 - (r_{ic} - c_{ic})^2\right)$ exactly, so VRMSE and relative L_2 are computed without reconstructing fields; captured $\leq U$ held on every cell. On the heat plume, the "state" is the training-mean temperature field and the "change" is the permeability input, since the task is not autoregressive; its study metric is VRMSE of the normalized temperature, not the benchmark's MAE.

Strategies and corrector. A *tile* is one Haar scale, orientation and position, all channels together. *Pooled*: global statistics in, one broadcast constant per channel out. *Static*: the K tiles with the largest mean training error. *Proxy*: top-K by the energy of the backbone's own predicted change. *Learned*: a scorer trained to rank residual energy. *Gain*: a scorer trained on the co-processor's own realized per-tile gain. *Random*. Atoms sit at their tiles (x, y from position, z from scale); detunings come from tile features and global context. Each tile receives $c_i = s_i(W_{\kappa(i)} f_i + A\, o_i)$ with a per-(scale, orientation) linear map W shared by every arm, the core's readout $o_i = (\langle Z_i \rangle, \langle X_i \rangle, \langle Y_i \rangle)$ and a zero-initialized A. The amendments: global context for tile co-processors after the first cells; planetswe excluded because its released checkpoints do not reproduce.

Co-processor. Tile features are scale-free (coefficients divided by the tile's own scale s_i, clipped), plus log scale, scale and orientation one-hots, normalized position and the window's global statistics. Registers are groups of 10 atoms in Morton (Z-order) order of their tiles; each atom's detuning per layer is $b_{\ell i} + \pi \tanh(w_\ell^\top f_i)$ (weights shared across atoms), positions are tile coordinates divided by learned per-axis blockade scales plus a learned scale-plane separation, and four layers of the propagator of §3 are applied. The DeepSets MLP sees $[f_i, \bar{f}]$ and has the nearest width to the register's parameter count (e.g. 316 vs 271 on TRL2D, 156 vs 165 on the heat plume). Scorers are two-layer MLPs trained with a listwise loss over each window's proxy shortlist, on residual energy (learned) or on the proxy-trained co-processor's realized per-tile gain (gain). AdamW, learning rate $3{\times}10^{-3}$, batch 64 windows, up to 40 epochs with patience 6 on validation; the loss is the negative captured share. A learning-rate check on TRL2D (10^{-3} to 10^{-4}) left the TRL2D null unchanged.

Analog–digital probe. Groups of 10 atoms with features $f_i \sim U[-1, 1]^4$ on random sites of a $4{\times}4{\times}2$ grid; targets per atom $\sin(\omega\, f_i^\top u)\cos(\omega\, \overline{f^\top v})$ (analog), $\mathrm{sign}(f_i^\top u)\,\mathrm{sign}(\overline{f^\top v})$ (digital);

20k/5k/10k groups, Adam 3×10^{-3}, 3,000 steps, best-validation checkpoint, test R^2, three seeds. Registered rule: an analog edge requires the register to beat its matched MLP on 3/3 seeds at ≥ 3 of the five ω. Test R^2, mean of three seeds (*exploratory arms added after the $\omega \leq 4$ results; MLP: nearest width to the register's 84 parameters, 75; MLP$^+$: next width up, 93):

core	ω=0.5	ω=1	ω=2	ω=4	ω=8	ω=16	digital
register	0.999	0.998	0.990	0.873	0.000	-0.000	0.067
mean-field	0.803	0.971	0.990	0.642	-0.000	—	0.226
MLP	1.000	0.999	0.979	0.953	-0.000	-0.000	0.719
MLP$^+$	1.000	1.000	0.989	0.956	-0.000	-0.000	0.920
MLP×4	1.000	1.000	0.986	0.989	0.012	-0.000	0.954
register, linear enc.*	—	—	—	0.874	-0.000	-0.000	—
sine-MLP*	—	—	—	0.988	-0.000	-0.000	—

Outside physics, on 22 TDC ADMET endpoints the register trails its matched MLP on 13 of 22 (it leads on 9).

Positive control: quantum-generated targets. Registered before any run, in the same harness: each atom's target is $\langle Z_i \rangle$ after a 30-step Trotterized interacting Rydberg quench (constant drive Ω=1.5, input-set detunings δ_i=$2f_i^\top u$, V_{ij}=r_{ij}^{-6} with nearest neighbours at $V\approx$1.8). The teacher's depth and fixed geometry lie outside the trained 4-layer ansatz. After a design critique, and before any register result, we added message-passing cores that see the teacher's exact pair interactions. Test R^2 (three seeds): mean-field register 0.671, register 0.433, GNN 0.333 and 0.357 (4×), MLPs 0.19–0.20. The register beats every MLP and GNN on every seed, so the harness separates cores sharply when the target calls for it. It loses to its mean-field twin on every seed, however, so the registered rule is not met. What helps on these targets is the Rydberg structure (geometry and r^{-6} interactions), which the classical mean-field model carries. At ten atoms and this training budget, entanglement adds nothing.

E Compute cost

All runs used one NVIDIA DGX Spark (GB10 Grace Blackwell chip, 128 GB of unified CPU–GPU memory, Arm64 host); no cluster or cloud compute was used. Median wall-clock per training step on the TRL2D neck insertion (released CNextU-net, $4 \times 16 \times 128 \times 384$ input, forward, backward and AdamW in float32, 25 steps after 5 warm-up, two repetitions, GPU shared with other jobs): backbone 504 ms, matched MLP 511 ms, 30-atom register 595 ms. The exact statevector propagator is the only added cost; inference with a trained register can use its product-state surrogate at the MLP's cost because the trained register is inference-inert.

Qubits are not parameters. Every register–MLP comparison is at matched trainable parameters, never at one parameter per qubit. A 10-atom group evolves a $2^{10} = 1{,}024$-amplitude state, but only its detuning maps, drives, evolution times and coordinates are trained, and it returns three expectation values per atom, not its amplitudes. In the routed co-processor on TRL2D at 30 qubits the register core has 244 trainable parameters and its matched MLP 199, which leaves the MLP two hidden units per atom. The qubit axis of Appendix N is therefore a different axis from the parameter counts of classical world models: 30 to 50 qubits add a few hundred trainable parameters to a backbone of 18.6 million (the TRL2D CNextU-net). That the register still ties an MLP with no more trainable parameters says these tasks, at these sizes, do not draw on its larger state space.

F Scores relative to published numbers (all classical lifts)

*Fusion Equilibrium: best development-leaderboard entry on the public DIII-D test fold, September 25, 2026 (0.9932 in August). Our TEST is a held-out split of the training shots, not the same test set. MHD_64's released checkpoint evaluates to 0.5447 per-channel VRMSE, not the published 0.1633 (our gated, fine-tuned backbone: 0.1725 on the full test split), and shear_flow's to 0.9946, not 0.8080; we make no comparison to the published numbers on those two rows. Persistence ($\hat{y}_{t+1} = x_t$, full official test splits, the benchmark's own mean VRMSE) is a necessary baseline on

Table 6: Our best numbers against published ones. Every gain is attributable to classical components (the register effect on the TRL2D long-horizon row did not replicate at four seeds, §5.2) (fine-tuning, correctors, ensembles, edge-weighted training); the register's contribution is measured in Tables 1–2. Ensembles are labelled as such.

Benchmark	Published best	Released checkpoint, our evaluator	Persistence	Ours (component)
TRL2D one-step VRMSE ↓	CNextU-net 0.1956	0.1934	0.5761	**0.1909** (gated fine-tuning, 3-backbone ensemble; plain 3-seed ensemble 0.1929)
TRL2D rollout 13:30 ↓	CNextU-net 1.01	1.0132	—	0.9528 (TD fine-tune with register, 3 seeds; 0.9660 with the registered fourth)
RealPDEBench foil rel. L_2 ↓	released U-Net	0.00948	—	0.00931 (linear corrector + 0.2 U-Net-real, VALID-weighted)
TokaMark 1-1 ↓	TokaMind-Scratch 0.140 (FT-Base 0.150)	—	n/a	0.1379 (TokaMind-Scratch recipe, our run, 1 seed; registered 4-member ensemble 0.1400)
rayleigh_benard ↓	TFNO 0.6566	0.6679	0.2024	**0.0403** (gated, metric-aligned fine-tuning; equal-weight 3-backbone ensemble; backbone alone 0.4563)
Fusion Equilibrium Ch. 1 S ↑	0.9959*	—	n/a	TEST 0.9910 (VALID-selected 7-member ensemble, VALID 0.9913; below the best)

the Well rows: it scores 0.2024 on rayleigh_benard and 0.0646 on shear_flow, below every published model and every backbone we trained (0.36–0.57 and 0.22–0.34), although the backbones win 84% and 54–63% of the individual windows; their mean is dominated by a heavy tail of low-variance windows. On TRL2D (0.5761) and MHD_64 (0.4183, beaten on every window) the backbones are far better than persistence. We therefore make no state-of-the-art statement from backbone numbers on rayleigh_benard or shear_flow. With the band gate and metric-aligned fine-tuning (Appendix G), rayleigh_benard reaches 0.0403 on the full test split, a fifth of persistence. This reflects metric alignment, not better convective dynamics: the convection-onset windows ($t_0 < 10$, 3.6% of test windows) carry 39% of the backbone's mean VRMSE (backbone 4.86, persistence 0.038 on them), and the classical gate leans on persistence per band and window where the backbone is unreliable. Recent foundation and compact models on The Well (McCabe et al., 2025; Sankaran et al., 2026) report the *median* one-step VRMSE over windows, not the benchmark's mean. Because the mean is dominated by the heavy tail discussed above, their numbers are not comparable to the mean-based scores in this table.

Where TokaMark 1-1 is lost. Training TokaMind from scratch with its published recipe gives 0.1379 in our evaluator (one seed), matching the reported 0.140; the registered equal-weight ensemble of this model with three fine-tuned FT-Base seeds gives 0.1400: the FT-Base members (0.150–0.151 each) hold three of the four equal weights and dilute the Scratch model. We report this registered ensemble as is and fit no weights after the fact. With the neck swapped at the same slot on the Scratch recipe (one seed each), the register neck gives 0.1404 and the MLP neck 0.1391: neither helps this backbone. The registered ensembles tie: 0.1400 (vanilla), 0.1400 (register neck), 0.1402 (MLP neck). Per target, our 3-seed register-neck FT-Base ensemble (0.1454 overall) beats TokaMind-Scratch (0.140; per-target values from its Table A.2 (Boschi et al., 2026)) on the plasma geometry: x-point z 0.0201 vs 0.022, x-point r 0.1654 vs 0.170, magnetic-axis z 0.0725 vs 0.075. The whole gap is in the three pressure ratios: β_{pol} 0.1655 vs 0.153, β_N 0.1687 vs 0.157, β_{tor} 0.1680 vs 0.157. The vanilla and MLP-neck ensembles show the same pattern. By shot length, the mean per-shot NRMSE is 0.172 on the shortest quarter of test shots ($\leq$218 windows, mostly early terminations) against 0.131–0.141 on the rest, so the gap is concentrated in short, disrupted discharges. It is not a learnable bias. On the test windows, the share of the β squared error explained by its mean within bins of shot length $\times$ time in shot is at most 0.6% (an in-sample upper bound for any additive correction on those bins), and the reconstructed β labels jump 1.5–9.6$\times$ more than the predictions between consecutive windows. We read this as noise in the EFIT labels near termination and fit no correction; the rule was fixed before the numbers.

G Evaluation-error analysis: persistence, a frequency-aware gate, and metric-aligned fine-tuning

Where the Well backbones fail. The benchmark's score is the mean over windows of a per-window, per-field VRMSE, which divides each window's error by that window's variance. On rayleigh_benard and shear_flow this mean is dominated by a heavy tail of low-variance windows (early convection onset; slowly evolving shear layers). There the backbones, trained with plain MSE on normalized fields, give those windows almost no weight. Persistence ($\hat{y}_{t+1} = x_t$) is therefore a strong baseline: on the full official test splits it scores 0.2024 and 0.0646, below the published leaders (0.6566, 0.8080) and below every backbone we fine-tuned, although the backbones win 84% and 54–63% of individual windows. A per-window oracle choice between backbone and persistence would score 0.044–0.048 and 0.036–0.039. On TRL2D (0.5761) and MHD_64 (0.4183) persistence is far behind.

A classical, frequency-aware gate. A related classical design in Hybrid Neural World Models (Lakshmanan & Chopra, 2026) gates a neural surrogate against a reference-solver fallback using an error map that detects sharp dynamical events (shocks, fronts). In our architecture, the fallback is persistence, evaluated per window and Fourier octave; rather than abandoning surrogate throughput via an expensive numerical solver fallback, QUOTA routes the remaining frequency-localized residual modes directly to an agile quantum (or matched classical) co-processor. We split the backbone's increment over persistence into five dyadic ℓ_∞ Fourier shells P_b (orthogonal projectors summing to the identity) and predict $\hat{y} = x_t + \sum_b \alpha_{c,b} P_b\big(p(x) - x_t\big)$ per field c. With α fitted in closed form on training windows (the *blend*) or predicted per window from causal features (the *gate*: band energies of $p - x_t$ and of the last input increment, normalized by the last input frame's variance), the gated residual's energy and every mode coefficient are exact functions of cached band statistics (`gate_energy_exact`, `gate_coeff_exact`; the per-band oracle gate is no worse than persistence or the backbone, `gate_opt_le`). The blend keeps the backbone's low bands and falls back to persistence for the high bands of the velocity fields ($\alpha \approx 1$ at low k, 0–0.1 at the top octave). A first version normalized the gate's features by the *target* frame's variance. We caught this leak, archived those runs, and reran with causal features. Every number below is causal; the change moved results by less than 1%.

The routed register after the gate. The residual that remains is written in band-wise POD modes (POD computed inside each band, so each qubit carries a signal of known band, and per band $\text{captured}_b \le \text{span}_b \le \text{full}_b$ holds for every window, `band_span_le_energy`). A learned router sends 30 modes per window to the register, and a classical head takes the rest. Without supervision, the router left the register nearly idle: it put 92% of its slots in the top octave, which holds almost none of the reachable error. Supervising the router toward each window's largest reachable error moved 80% of the slots to the lowest band. Table 7 reports the matched cores (register, GNN on the same learned graph, MLP) with the supervised router. With supervision, routing matches the allocation thesis: the qubit-routed modes carry most of the co-processor's capture, in the band where the reachable error is (Table 8). But the matched classical cores, given the same routed modes, capture as much: the cores tie. We attribute the gains to the gate, the routing and the classical head.

Table 7: Evaluation-error analysis on the Well rows (VRMSE, lower is better). Backbone: the fine-tuned (or released) CNextU-net. Blend: TRAIN-fitted band blend with persistence. Gate: per-window causal band gate. The last three columns stack the routed co-processor (30 atoms or matched classical cores) on the gate; mean $\pm$ SD over seeds. Full official test splits except † (stratified test sample, every trajectory $\times$ 8 time bins, 4 windows each; validated against full-split per-window scores to within 2%). ‡Claimed; the released checkpoint evaluates to 0.5447.

Row (backbone)	Published	Persistence	Backbone	Blend	Gate	+Register	+GNN	+MLP
shear_flow (register neck)	0.8080	0.0646	0.2297	0.0451	0.0272	0.0201 ± 0.0002	0.0205 ± 0.0002	0.0200 ± 0.0003
rayleigh_benard† (register neck)	0.6566	0.2024	0.4321	0.1420	0.0420	0.0413	0.0415 ± 0.0003	0.0416 ± 0.0004
rayleigh_benard† (vanilla)	0.6566	0.2024	0.4252	0.1438	0.0436	—	—	—
TRL2D (register neck)	0.1956	0.5761	0.1931	0.1924	0.1921	0.1922	0.1922 ± 0.0001	0.1922
TRL2D (released)	0.1956	0.5761	0.1934	0.1926	0.1924	0.1923 ± 0.0001	0.1923 ± 0.0001	0.1923 ± 0.0001
MHD_64 (vanilla)	$0.1633^\ddagger$	0.4183	0.1771	0.1761	0.1735	0.1733	0.1733	0.1733

Metric-aligned end-to-end fine-tuning. The fix belongs in training, not only in stacking. We fine-tune the whole stack end to end: backbone (vanilla, MLP neck, or register neck), gate, and the benchmark metric itself as the loss. Selection uses a stratified validation sample; the selected configuration is scored once on the full test split (Table 9).

Three readings. First, the closed-form gate on the frozen backbone already captures the gain. After exact gate pre-training, the joint stage improves VALID by at most 0.001, and on shear_flow the best checkpoint is the starting one, so end-to-end retraining is not needed. Second, the loss matters only together with the gate. On rayleigh_benard, MSE with the gate gives 0.1415, VRMSE without the gate 0.4331, and both together 0.0425–0.0742. The VRMSE loss gives low-variance onset windows the gradient that MSE takes away; the credit goes to that classical loss, not to the neck. Third, the three backbones tie within one seed (TRL2D 0.1921–0.1922, MHD_64 0.1727–0.1728), and no register edge is claimed from the single-seed differences on rayleigh_benard and shear_flow. The registered equal-weight ensemble of the three backbones (E-ft; registered before any ensemble score)

Table 8: Where the register's work goes, per Fourier band, after the gate (route-supervised router, seed 1). Shares of each window's error before the gate, averaged over test windows. full_b: error in band b after the gate. span_b: part of it on the band's POD modes. Then what the qubit-routed modes and the classical head capture, and the share of the 30 qubit slots that land in the band. The chain $\text{captured}_b \leq \text{span}_b \leq \text{full}_b$ is checked on every window; the largest violations are fp16 rounding.

Row	Band	full_b	span_b	Captured (qubits)	Captured (classical)	Qubit slots
shear_flow	b_0	0.3891	0.3433	0.2167	0.0419	95%
	b_1	0.0444	0.0186	0.0050	0.0008	2%
	b_2	0.0360	0.0080	0.0005	0.0000	3%
	b_3	0.0297	0.0016	-0.0000	-0.0004	0%
	b_4	0.0342	0.0006	0.0000	-0.0002	0%
	max violation: captured≤span -1.5e-10, span≤full -5.3e-08					
rayleigh_benard	b_0	0.0476	0.0359	0.0186	0.0011	81%
	b_1	0.0180	0.0047	0.0008	0.0001	12%
	b_2	0.0151	0.0017	0.0009	0.0000	2%
	b_3	0.0191	0.0008	0.0007	-0.0000	1%
	b_4	0.0267	0.0008	0.0004	0.0000	3%
	max violation: captured≤span -3.2e-09, span≤full -6.0e-07					

Table 9: End-to-end fine-tuning with the band gate (seed 1). Loss: VRMSE (the benchmark metric) or MSE. Schedule: joint from the start, or exact gate pre-training on 3,000 frozen-backbone training windows, then joint. VALID: stratified sample (start → best). Test: stratified estimate, and the full split for the selected cells.

Row	Backbone	Loss	Head	Schedule	VALID start	VALID best	Test (strat.)	Test (full)
shear_flow	MLP neck	vrmse	gate	pre	0.0284	0.0284	0.0280	—
shear_flow	register neck	vrmse	gate	pre	0.0275	0.0275	0.0271	—
shear_flow	vanilla	vrmse	gate	pre	0.0301	0.0294	0.0298	—
rayleigh_benard	MLP neck	vrmse	gate	joint	0.1436	0.0681	0.0659	—
rayleigh_benard	MLP neck	vrmse	gate	pre	0.0454	0.0437	0.0438	—
rayleigh_benard	register neck	mse	gate	joint	0.1426	0.1426	0.1415	—
rayleigh_benard	register neck	vrmse	gate	joint	0.1426	0.0777	0.0742	—
rayleigh_benard	register neck	vrmse	gate	pre	0.0422	0.0422	0.0425	—
rayleigh_benard	register neck	vrmse	plain	joint	0.4470	0.4470	0.4331	—
rayleigh_benard	vanilla	vrmse	gate	joint	0.1441	0.0891	0.0854	—
rayleigh_benard	vanilla	vrmse	gate	pre	0.0438	0.0438	0.0441	—
TRL2D	MLP neck	vrmse	gate	pre	0.1930	0.1918	0.1922	—
TRL2D	register neck	vrmse	gate	pre	0.1928	0.1919	0.1921	—
TRL2D	vanilla	vrmse	gate	pre	0.1926	0.1916	0.1921	—
MHD_64	register neck	vrmse	gate	pre	0.1723	0.1714	0.1728	—
MHD_64	vanilla	vrmse	gate	pre	0.1722	0.1714	0.1727	—
shear_flow	E-ft (3 backbones, equal)	vrmse	gate	pre	—	0.0256	—	0.0252
rayleigh_benard	E-ft (3 backbones, equal)	vrmse	gate	pre	—	0.0400	—	0.0403
TRL2D	E-ft (3 backbones, equal)	vrmse	gate	pre	—	0.1917	—	0.1909
MHD_64	E-ft (2 backbones, equal)	vrmse	gate	pre	—	0.1726	—	0.1725 (vanilla)

is chosen on VALID over the best single backbone on rayleigh_benard (0.0400 vs 0.0422) and scores 0.0403 on the full test split. The members' residual fields correlate at 0.59–0.64, symmetrically across pairs, so the gain is ordinary variance reduction; the register neck is one member and gets no credit for it. On MHD_64 the two members' residuals correlate at 0.999, the ensemble ties them on VALID, and the selected single backbone scores 0.1725 on the full test split. On shear_flow (residual correlations 0.49–0.68) E-ft is selected and scores 0.0252 on the full test split. On TRL2D it is selected by 0.0001 on VALID and scores 0.1909 on the full test split, below the published 0.1956. There the members' residuals correlate at 0.997–0.999, which bounds the ensembling gain at about 0.0001, so the step from the frozen-backbone gate (0.1921, Table 7) to 0.1909 comes from the gated, metric-aligned fine-tuning itself.

What is left: a transform diagnosis. We decomposed the final models' test error per Fourier octave, per Haar wavelet level and in the z-domain (Figure 9; nothing fitted). With $D = y - x_t$ the true change and $P = \hat{y} - x_t$ the predicted change, the magnitude-squared coherence of P and D in an octave gives its Wiener floor, the least error any spatially uniform linear gain on P in that octave can reach. On rayleigh_benard the models sit on this floor in every octave and field; on TRL2D they are within 0.06 of it (in units of the octave's true-change energy); adding the last observed change $x_t - x_{t-1}$ as a second regressor lowers the floor by at most 0.03. No spatially uniform linear filter per

octave can remove what is left; nonlinear, cross-scale or local models might. In the z-domain, the lag-one pole of each octave's change matches the truth within 0.02 on TRL2D and rayleigh_benard; in the top octaves both sit near $-\frac{1}{2}$, the value that frame-to-frame white content gives ($x_t = s_t + e_t$ with e i.i.d. and s smooth yields $\mathrm{corr}(D_{t+1}, D_t) = -\frac{1}{2}$). On shear_flow the models' pole there is near 0 against the truth's -0.43 to -0.51, the one mismatch we found. Wavelets locate the error: on TRL2D 98% of its energy lies in 5% of the Haar coefficients, along the mixing interface, and pressure carries most of it; on rayleigh_benard the worst windows are convection onset (t_0 11–30), where a thin, grid-scale layer at one wall starts to move, 62% of the buoyancy error sits at the finest level, and the octave-4 floor is 0.34–0.50. A deterministic model's least-squares answer to such unpredictable content is to shrink it, which the amplitude ratios (0.72–0.88 in the top octave) show; probabilistic prediction is the principled alternative (Appendix Q). A registered two-tap filter per octave on top of the final models ($\alpha P + \beta(x_t - x_{t-1})$, fitted on training windows) improves TRL2D's VALID score by 0.12%, below its 0.5% bar, and is dropped. On shear_flow, where the pole mismatch is, it improves VALID by 16% and the stratified test sample from 0.0252 to 0.0202 (72% of windows lower; scored once). Its fitted coefficients say where the model failed: the large-scale pressure change is better predicted by continuing the last observed change ($\beta = 0.72$) than by the model's own change ($\alpha = 0.31$), and the top pressure octaves fall back to persistence with $\beta \approx -\frac{1}{2}$, the white-content reversion. The gain is metric alignment again: on a subsample of the test windows it comes from the three lower quintiles of target variance ($+29$ to $+33\%$), while the two upper quintiles get worse (-15%, -30%), so the filter corrects quiet windows, not active flow. It is classical and one-step, acts on velocity components separately (incompressibility is not enforced), and its behaviour inside a rollout is untested. Finally, against the released checkpoint on the same 873 TRL2D test windows, the final model is lower on 6 of 9 test trajectories (the other three within 0.0006): paired Wilcoxon over trajectories $p = 0.049$, trajectory-bootstrap 95% CI of the mean difference $[-0.0046, -0.0006]$.

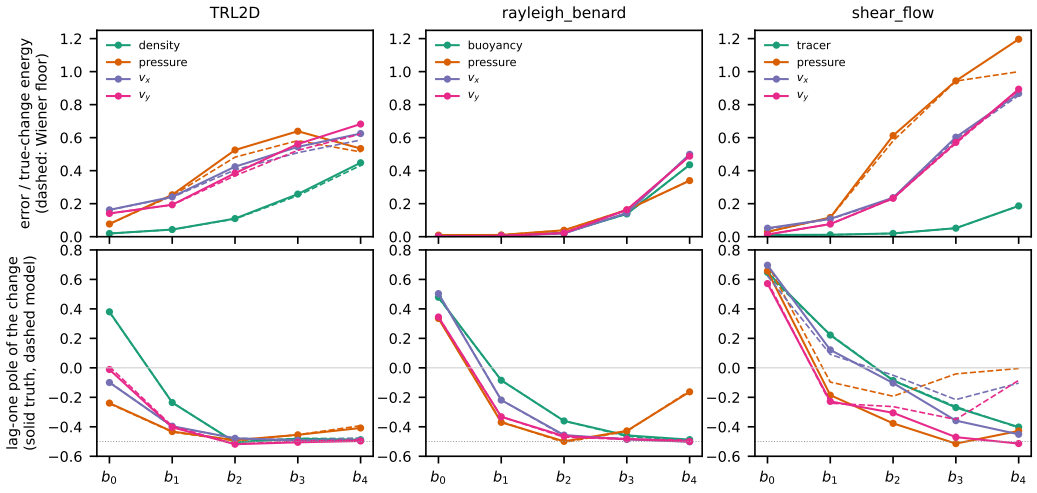

Figure 9: What is left after the gate, per Fourier octave b_0 (coarsest) to b_4. Top: error energy over the octave's true-change energy (solid) and its Wiener floor, $1-$coherence (dashed). Bottom: lag-one pole of the octave's change, truth (solid) against the final model (dashed); the dotted line marks $-\frac{1}{2}$, the value frame-to-frame white content gives. TRL2D: full test split; rayleigh_benard and shear_flow: stratified test samples.

H Per-seed results for the TRL2D official and sliding-start protocols

Four-seed replication. Registered before any run of seed 1340: one more seed per arm, with the same recipe, checkpoint selection (best rollout VALID) and evaluation. The effect counts as replicated only if seed 1340's primary test is significant against both fine-tuned controls, Holm across four seeds. It is not. At seed 1340 all three arms selected checkpoints from the first 72–84

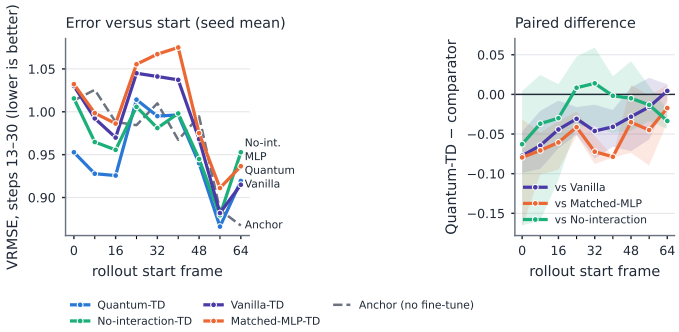

Figure 10: Rollout error versus start frame under the sliding-start protocol. *Left:* seed-mean 13:30 VRMSE per arm at each start (anchor dashed). *Right:* paired difference, register minus comparator (seed mean, band = seed range; below zero = register better). The edge over the fine-tuned classical controls persists across starts; the difference from the anchor is negative only at early starts.

Table 10: TRL2D long-horizon window 13:30 test VRMSE (TD fine-tuning, 3 seeds). Right: seeds on which the register beats that model, by the pre-registered test over start frames (Holm-corrected) and by an uncorrected test over all 81 trajectory-start pairs. Per-seed values and bootstrap intervals: Appendix H.

Model	from the official start frame	mean over 9 start frames	seeds better (Holm)	seeds better (81 pairs)
+ 30-atom register	$\mathbf{0.9528 \pm 0.0542}$	$\mathbf{0.9486 \pm 0.0231}$	—	—
+ register, V_{ij}=0 (no interactions)	1.0156 ± 0.0433	0.9664 ± 0.0264	1/3	1/3
released checkpoint (anchor)	1.0132	0.9707	0/3	1/3
+ register, mean-field (Gutzwiller)	1.0495 ± 0.1219	0.9782 ± 0.0418	0/3	0/3
vanilla continued	1.0301 ± 0.0449	0.9867 ± 0.0196	**2/3**	**3/3**
+ matched MLP	1.0322 ± 0.0259	1.0041 ± 0.0160	**2/3**	**3/3**

steps (seeds 1337–1339: steps 5,726–9,309), and the register is the worst of the three: start-averaged 13:30 VRMSE 0.9568 against 0.8996 (MLP) and 0.8939 (vanilla), worse on all nine trajectories against each. Over four seeds the register is significant against the MLP on 2/4 seeds and against vanilla on 1/4 (Table 11). The four-seed start-averaged means are 0.9506 (register), 0.9635 (vanilla) and 0.9780 (MLP), and the per-seed tests support no edge. A secondary analysis scores every arm's final step-10,000 weights with the same tests. It was registered after the early selection was seen for the register and before any final-step evaluation, and it cannot change the primary verdict. There the register is significant against the MLP on 0/4 seeds and against vanilla on 2/4, so the edge over the MLP seen on three seeds depended on validation-based checkpoint selection.

Table 11: TRL2D TD replication, four seeds (1337–1340): 13:30 test VRMSE (mean ± SD over seeds) at the VALID-selected checkpoints (primary) and at the final training step (secondary), the seeds on which the register is significantly better (per-trajectory Wilcoxon on start-averaged error, Holm across four seeds), and the seed-1340 official-start value.

Checkpoint	arm	official start	all 9 starts	reg. better (Holm, 4 seeds)	seed 1340
VALID-selected	register	0.9660 ± 0.0515	0.9506 ± 0.0193	—	1.0056
	matched MLP	1.0101 ± 0.0490	0.9780 ± 0.0539	2/4	0.9439
	vanilla continued	0.9892 ± 0.0897	0.9635 ± 0.0491	1/4	0.8664
final step	register	1.0467 ± 0.0684	1.0242 ± 0.0203	—	1.1009
	matched MLP	1.0109 ± 0.0507	0.9723 ± 0.0373	0/4	0.9841
	vanilla continued	1.1236 ± 0.1669	1.0877 ± 0.0978	2/4	1.1456

Table 12 lists every TD arm's per-seed official-start (start-0) 13:30 test VRMSE, its all-start value, and its rollout-validation selection best (VRMSE@8 at the selected checkpoint). Table 13 gives the pre-specified statistics of the register against each comparator. The sliding-start protocol reproduces

the official protocol exactly at start 0 (checked to machine precision). Its primary test is a per-trajectory paired Wilcoxon on 13:30 VRMSE averaged over the nine starts (n=9, one-sided, Holm across seeds); the secondary uses per-(trajectory, start) pairs (n=81, correlated within trajectory, uncorrected). Per-window p-values on the one-step endpoint are anti-conservative because windows overlap within trajectories and are read together with seed-level consistency.

Arm	13:30, official start (s1337/s1338/s1339)	13:30, all starts	selection best
official checkpoint (anchor)	1.0132	0.9707	—
vanilla continued	1.0787 / 1.0216 / 0.9901	1.0091 / 0.9784 / 0.9727	0.5846 / 0.5909 / 0.5750
+ matched MLP	1.0328 / 1.0060 / 1.0578	0.9993 / 1.0221 / 0.9911	0.5716 / 0.5796 / 0.5802
+ 30q, V_{ij}=0	1.0202 / 0.9701 / 1.0564	0.9437 / 0.9602 / 0.9954	0.5923 / 0.5807 / 0.5739
+ 30q, mean-field	1.1882 / 0.9592 / 1.0011	1.0222 / 0.9732 / 0.9390	0.5847 / 0.5870 / 0.5838
+ 30q Rydberg	0.9927 / 0.9745 / 0.8911	0.9620 / 0.9618 / 0.9219	0.5783 / 0.5771 / 0.5642

Table 12: Per-seed values behind Table 10. The mean-field arms' official-start 6:12 values are 0.5029/0.5221/0.5397; all arms' 6:12 means are 0.5216–0.5278 (anchor 0.5482).

comparator	seed	mean diff	prim. win	prim. p	Holm	sec. win	sec. p	bootstrap CI_{95}
vanilla	1337	-0.0471	0.78	0.0195	0.0391*	0.59	5.0×10^{-3}	$[-0.0833, -0.0129]$
vanilla	1338	-0.0167	0.78	0.1016	0.1016	0.57	2.3×10^{-2}	$[-0.0393, +0.0101]$
vanilla	1339	-0.0508	0.89	0.0039	0.0117*	0.77	7.5×10^{-8}	$[-0.0702, -0.0260]$
matched MLP	1337	-0.0372	0.78	0.1016	0.1016	0.53	1.9×10^{-2}	$[-0.0795, +0.0079]$
matched MLP	1338	-0.0603	1.00	0.0020	0.0059*	0.75	5.2×10^{-8}	$[-0.0889, -0.0339]$
matched MLP	1339	-0.0692	1.00	0.0020	0.0059*	0.77	1.0×10^{-8}	$[-0.0929, -0.0462]$
anchor	1337	-0.0086	0.56	0.5449	1.0000	0.44	4.9×10^{-1}	$[-0.0981, +0.0725]$
anchor	1338	-0.0089	0.44	0.6328	1.0000	0.48	5.0×10^{-1}	$[-0.0900, +0.0571]$
anchor	1339	-0.0488	0.56	0.1504	0.4512	0.56	3.5×10^{-2}	$[-0.1206, +0.0249]$
V_{ij}=0	1337	$+0.0183$	0.11	0.9512	1.0000	0.30	1.0	$[-0.0706, +0.0737]$
V_{ij}=0	1338	$+0.0016$	0.44	0.5000	1.0000	0.54	4.6×10^{-1}	$[-0.0333, +0.0391]$
V_{ij}=0	1339	-0.0735	0.89	0.0039	0.0117*	0.67	9.5×10^{-5}	$[-0.1152, -0.0350]$
mean-field	1337	-0.0602	0.44	0.4102	0.5703	0.40	5.4×10^{-1}	$[-0.2393, +0.0877]$
mean-field	1338	-0.0115	0.44	0.2852	0.5703	0.53	2.0×10^{-1}	$[-0.0349, +0.0090]$
mean-field	1339	-0.0172	0.56	0.1504	0.4512	0.51	1.2×10^{-1}	$[-0.0447, +0.0101]$

Table 13: Sliding-start protocol: register minus comparator on start-averaged 13:30 VRMSE, per seed. *: Holm-adjusted $p < 0.05$. CI: trajectory cluster bootstrap (10,000 resamples) of the mean paired difference.

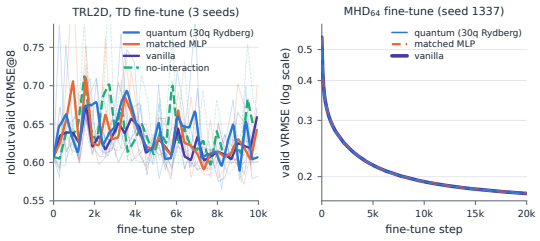

Figure 11: Fine-tuning dynamics. *Left:* TRL2D TD fine-tuning from the official checkpoint (thin: seeds; bold: 3-seed mean of the rollout-validation selection metric); arms overlap step by step, and the test-rollout separation is not visible in validation curves. *Right:* `MHD_64` (seed 1337, log scale); all arms converge to per-channel validation VRMSE ≈0.177.

I QUBIT-LADDER DETAILS

Per-seed VRMSE@8 (s1337/s1338/s1339) under TD: 5 atoms 0.9245/0.8053/0.7982; 10 atoms 1.1810/0.7799/0.8387; 20 atoms 1.2486/0.7843/0.7478; 30 atoms 0.903/0.826/0.806; matched MLP 1.098/0.933/0.829; mean-field pilot 0.891/0.825/0.802. Seed 1337 is the outlier at every interior count. Under the one-step arena objective the means were 30 atoms 1.1002 ± 0.1939, matched MLP 1.0574 ± 0.0574 and bare backbone 0.9637 ± 0.0048: the adapters' edge appears only under TD. A smoke-scale CMA-ES configuration search (about 200 evaluations per regime) was not competitive with gradient training.

J MHD_64

The released `MHD_64` checkpoint evaluates to 0.5447 (test) / 0.5464 (validation) per-channel VRMSE under our pipeline (≈0.66 field-grouped, the convention of the published 0.1633); the same pipeline reproduces the TRL2D anchor to 1.1%, which points to, but does not prove, a discrepancy in the released artifact. Four of the nine 20k-step runs stopped at about 7.5k steps; the primary row uses the two seeds at which all three arms reached the same length (backbone 0.1897 ± 0.0178, MLP 0.1897 ± 0.0178, register 0.1896 ± 0.0177; the arms agree to four decimals and the gate never opens). All arms beat the released artifact on every one of 960 test windows.

K LEARNED REGISTER CONFIGURATIONS

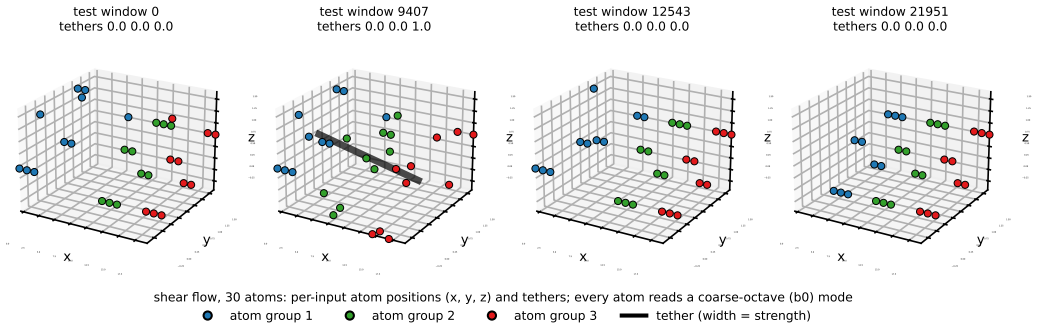

Figure 12: Per-input qubit configuration in 3-D (routed register, 3-D variant of Appendix N, shear flow, 30 atoms, seed 1): atom positions (x, y, z) of the three 10-atom groups (drawn side by side) on four test windows, and tether strength between groups (line width). Positions move with each input (mean 0.33, at most 1.36 grid units from the mean layout; z spans 2.1), and on window 9407 the network switches on the tether between groups 1 and 3.

L VISUAL ROLLOUT FIELDS: TRL2D AND MHD_64

M DYNAMIC CONFIGURATION AND HARDWARE MAPPING

Table 14 states what a deployment would need. The neck register's 2-D geometry and global drives map directly, and distances are clamped from below as both devices require. Three features do not map as trained. Signed, per-layer, input-dependent local detunings would have to be re-encoded as a global detuning plus one static, one-signed weight pattern per input, which means one device program per input. $\langle X \rangle$ and $\langle Y \rangle$ readouts need a basis-rotating pulse before measurement. The configuration network also outputs a z offset, but it stays at zero in training (a planar start gives it no gradient), so every learned layout is 2-D.

Finite-shot readout. A device estimates each atom's $\langle Z \rangle$, $\langle X \rangle$ and $\langle Y \rangle$ from N shots per measurement setting. We re-evaluated the route-supervised co-processor register (3 seeds per row, trained

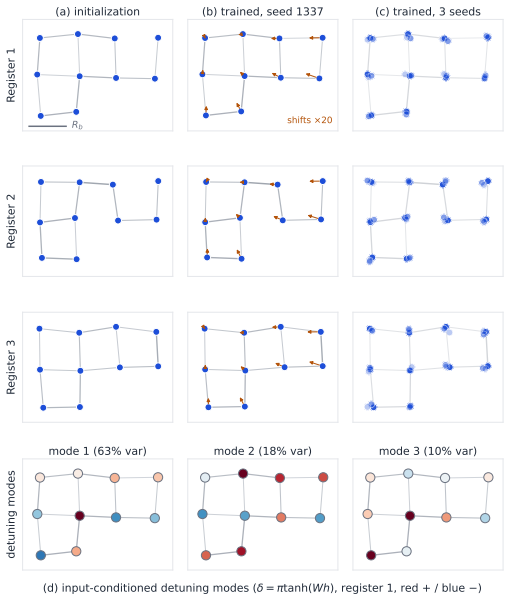

Figure 13: Register configurations (three blockade-decoupled 10-atom registers, trainable 2-D geometry, blockade-radius units). *(a)* Initialization (lattice with σ=0.05 R_b jitter). *(b)* TD-trained geometry (seed 1337, displacements ×20): a small coherent contraction (mean radial shift $-0.007\,R_b$). *(c)* Three seeds agree to $\leq 0.11\,R_b$ rms. *(d)* Per-input reconfiguration via detunings: the top three singular modes of the conditioning map hold 91% of its variance.

Table 14: What maps to current neutral-atom devices, from the vendors' public programming interfaces (QuEra Aquila via Amazon Braket's analog Hamiltonian simulation schema; Pasqal Fresnel and Pulser's detuning map modulator, DMM).

	Our registers	QuEra Aquila	Pasqal Fresnel
Geometry	2-D, 3 × 10 atoms; routed co-processor adds per-input 2-D offsets	user-defined 2-D arrangement	2-D on calibrated trap layouts; ≤80 atoms, ≥5 µm apart, within 38 µm
Drive	global Ω, ϕ per layer	global $\Omega(t)$, $\phi(t)$	global Rydberg channel
Detuning	signed, per atom, per layer, input-dependent	global $\Delta(t)$; local $\Delta_\ell(t)\,h_k$, $h_k \in [0,1]$ static (experimental)	global $\Delta(t)$; DMM: negative only, one waveform, static weights in $[0,1]$
Readout	$\langle Z\rangle$, $\langle X\rangle$, $\langle Y\rangle$	ground/Rydberg bitstrings	ground/Rydberg bitstrings

exactly as before) with each expectation replaced by $2\,\mathrm{Bin}(N,(1+\langle P\rangle)/2)/N - 1$; nothing was retrained with noise, and the rule was registered first: the register counts as shot-robust at N if the mean increase in test VRMSE stays below half its gap to the matched MLP. The increase is 7.6×10^{-5}, 8×10^{-6} and 1×10^{-6} on shear_flow and 3.2×10^{-4}, 3.4×10^{-5} and 4×10^{-6} on rayleigh_benard at $N = 100$, 1,000 and 10,000 (thresholds 5.5×10^{-5} and 1.8×10^{-4}). The register is shot-robust from $N = 1{,}000$ on both rows and not at $N = 100$. Because that gap is itself not significant, we also compare with seed spread: at $N = 1{,}000$ the increase is below the seed standard deviation of every core on shear_flow (1.8–3.4×10^{-4}) and of the classical cores on rayleigh_benard (3.1–3.7×10^{-4}), though not below the register's own spread there (2.2×10^{-5}). The measured cost per input is then $3N$ shots (three settings) when the tethers between the three 10-atom groups are inactive, and up to $9N$ when a learned tether makes the groups sequential, which it does on two of three shear_flow seeds. The matched classical cores need none of this. Calibration errors were tested the same way (registered first; 3 seeds, 3 repeats): Gaussian in-plane position jitter of 1% and 5% of the lattice spacing and a multiplicative detuning error of 1% and 5% per atom and layer. Jitter and the 1% detuning error raise test VRMSE by at most 7×10^{-6} on both rows; a 5% detuning error costs 1.4×10^{-4} on rayleigh_benard (within its threshold) and 7.3×10^{-5} on shear_flow (above its

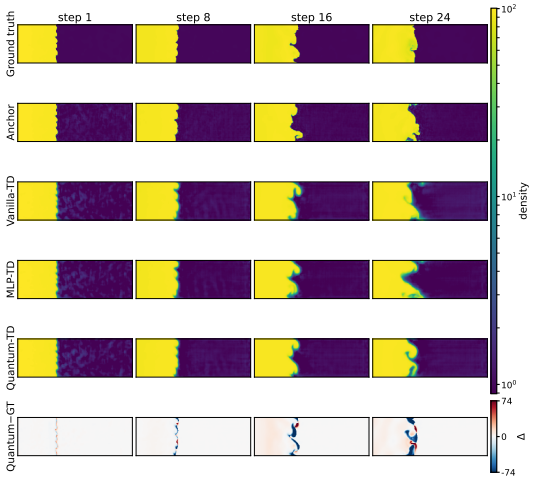

Figure 14: Ground truth vs model rollouts on the first TRL2D test trajectory (density, log scale; bottom: signed error of the TD register arm). Error concentrates at the mixing front. On this trajectory the anchor has the lowest late-step VRMSE (0.640/0.659 at steps 16/24 vs 0.890/0.712 for the register arm); headline claims rest on full-test-set averages.

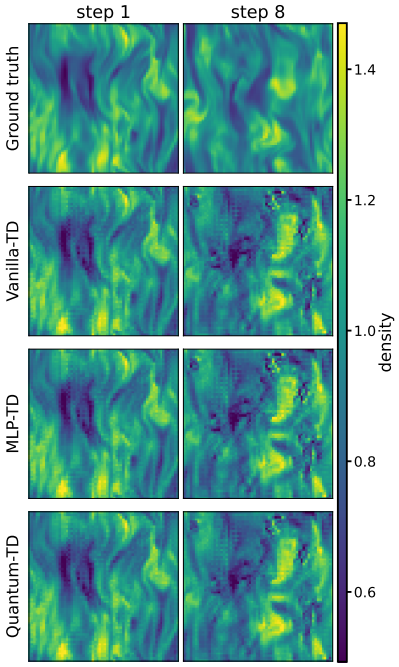

Figure 15: Ground truth vs model rollouts for 3D magnetohydrodynamics (MHD_64, valid trajectory 0, mid-z density slice, steps 1 and 8). Top to bottom: Ground truth, Vanilla-TD, matched MLP-TD, and Quantum-TD. All arms preserve coarse density topology; fine discrepancies concentrate in turbulent magnetic shear interfaces.

5.5×10^{-5} threshold). The register tolerates placement errors well and needs detunings calibrated to about 1%.

Detunings depend on each input; geometry, drives and evolution times are shared learned parameters in the reported arms, and each forward pass resets the register. An arena-only variant with input-conditioned positions $r_i(h) = r_i^0 + 0.6 \tanh(g_\phi(h)_i)$ reached 0.8951 ± 0.1367, worse than the static layout. Mapping to a neutral-atom device would require: replacing signed per-layer local detunings with the platform's addressable detuning patterns; enforcing minimum atom separation (clamping regularizes interaction energies but does not enforce physical spacing); replacing the first-order digital–analog split by continuous pulses; and a noise model beyond finite-shot readout. Optical tweezers assemble user-defined 2-D arrangements at the 100-atom scale (Barredo et al., 2018), with coherent arrays reported at 6,100 atoms (Manetsch et al., 2025); the registers here use at most 30 atoms in blocks of 10. The trained neck registers of the audit are inference-inert, so those deployed models need none of this; the routed co-processor is not, and its readout cost is given above.

N HYBRID SCALING WITH THE QUBIT BUDGET

Registered before the sweep (a 5-qubit smoke run only tested the code). The classical part is fixed: backbone, gate and a classical head on every mode. The routed co-processor sets every register degree of freedom it can per input: atom positions (x, y, z), detunings, which modes the atoms receive, and, new here, the strength of each tether between the 10-atom groups (a learned base plus a linear map of the routed modes' features). Drives and evolution times stay shared. Budgets

$K \in \{0, 5, 10, 20, 30\}$ ($K = 0$: classical part alone); register and parameter-matched MLP; three seeds on every row (full test on shear_flow, TRL2D and MHD_64, stratified test on rayleigh_benard). The backbone in the classical part is the vanilla CNextU-net, never trained with a register. A first version of this sweep used, on three rows, a backbone trained with a register in its neck; it is kept below as a check of how much the backbone matters.

Figure 5 (top) gives the hybrid's test VRMSE relative to its classical part. It falls with the budget on the two rows with predictable error: to 0.877 (register) and 0.900 (MLP) at 30 qubits on shear_flow, and to 0.949 and 0.954 on rayleigh_benard. It stays at 1.000 on TRL2D and MHD_64. Figure 6 shows where the added qubits work. The bottom row of Figure 5 explains the shape in error-energy space, where orthonormal modes add. The coverage $U(K)$, the share of the gated residual's energy in the best K modes, is computed from the test data for any K (no fitting) and bounds what K qubits can capture. It follows the closed form $U_\infty[1 - (1 + K/K_0)^{1-a}]$ within 0.015 on every row (0.008 or less except on TRL2D; U_∞ 0.416, 0.047, 0.555, 0.019 and a 1.92, 2.07, 1.85, 2.62 for shear_flow, rayleigh_benard, TRL2D, MHD_64), a power-law decay of mode energies. $U(K)$ saturates near 200 qubits because the band-wise POD basis spans only that share of the residual: the limit beyond it is the basis, not the qubit count, and a richer or input-adaptive basis (wavelet packets, a best basis per input) is what would raise it. The qubits reach $\eta(K) = \text{captured}/U(K)$ of the coverage, rising from 0.36 to 0.66 on shear_flow and from 0.42 to 0.66 on rayleigh_benard between 5 and 30 qubits; on TRL2D U_∞ is large but $\eta = 0$, the unpredictable residual of Appendix G. At 30 qubits the register captures 56% of what the basis allows on shear_flow (86% at its current efficiency); the shaded fans extrapolate beyond 30 qubits with η held at its 30-qubit value or at 1. The fan was tested at 50 qubits (five groups, same per-input layout), with ranges for the register's capture registered before any run: $[0.249, 0.380]$ on shear_flow and $[0.027, 0.040]$ on rayleigh_benard. The upper edges are bounds (capture cannot exceed coverage), so the test is whether η falls below its 30-qubit value. All six seeds fell inside (shear_flow 0.255, 0.253, 0.269; rayleigh_benard 0.029, 0.028, 0.028), near the lower edges, with η rising slightly (0.65 to 0.68 and 0.66 to 0.71), and the hybrid reached 0.869 and 0.940 of its classical part.

Where the gain comes from (registered, 3 seeds). The position head starts at zero and every atom at z=0, so the loss has no gradient in z and all trained layouts are 2-D. A control registered before it ran compares, at 30 atoms with the same routing, per-input 2-D positions with dynamic tethers (shear_flow 0.02073, rayleigh_benard 0.04261), a fixed grid without tethers (0.02071, 0.04258), and a bilayer grid that trains z (0.02110, 0.04271); the matched MLP scores 0.02129 and 0.04281. Neither "3-D helps" (1/3 seeds on each row) nor "geometry and tethers help" (2/3 and 0/3) passes its rule: the routing, not where the atoms sit, carries the gain. The matched MLP did as well or better (shear_flow 0.268, 0.263, 0.302; rayleigh_benard 0.028, 0.027, 0.028; hybrid 0.877 and 0.940), so at 50 qubits too the gain belongs to the routing, not the register. About 94% of the capture added between 30 and 50 qubits lands in the coarsest octave (shear_flow 0.222 to 0.248 of the error, rayleigh_benard 0.020 to 0.025); the finer octaves, whose error is unpredictable (Appendix G), barely move.

With the register-trained backbone the curves have the same shape: 0.858 (register) and 0.885 (MLP) at 30 qubits on shear_flow, 0.961 and 0.957 on rayleigh_benard, flat on TRL2D.

Exploratory, not the registered test: the register is below its matched MLP in 10 of 12 seed-by-budget cells on shear_flow (mean 1.8%; at 30 qubits 2 of 3 seeds, paired $p = 0.30$) and in 11 of 12 on rayleigh_benard (mean 0.5%; 3 of 3 at 30 qubits, $p = 0.19$); TRL2D and MHD_64 tie. With fixed tether strengths at 30 qubits the two tied (Appendix G). The registered rule requires a GNN arm and a majority of rows, so no quantum claim follows.

O A DISTRIBUTION OF CONFIGURATIONS PER INPUT

On TRL2D the qubits capture no predictable error at any budget ($\eta = 0$, Appendix N), and the change that the backbone misses behaves like frame-to-frame noise (Appendix G). A single configuration per input can only predict the mean of that noise. WeatherNext 3 (Rasp et al., 2026) builds its ensembles on functional generative networks (Alet et al., 2025): each member is drawn by feeding one low-dimensional noise vector into the network's conditional normalization layers, and training minimizes the fair CRPS over two members. We move the noise into the qubit configuration.

Method. For each window and ensemble member, a code $\xi \sim \mathcal{N}(0, I_{32})$ enters only the input-dependent configuration of Appendix N (30 qubits, band gate, supervised router, per-input tethers). It enters in three places:

- a feature-wise affine map (scale and shift) of the features that enter the core, the analogue of conditional normalization;
- the hidden layer of the network that sets each atom's x, y, z offset;
- the per-input tether logits.

Each input therefore gets a distribution over atom positions, detunings and tether strengths, and each draw is one ensemble member. The three maps start at zero, so training starts at the deterministic hybrid. The gate, the classical head and the routing stay deterministic, so any spread comes through the co-processor alone.

Loss. We minimize the fair energy score over two members, a strictly proper score for the whole field (Gneiting & Raftery, 2007). In the orthonormal band-wise POD basis it is exact and cheap. Per field, a member's distance to the target is $\sqrt{E_{\text{new},m}/D}$, which is that member's VRMSE. The distance between two members is $\|c_m - c_{m'}\|/\sqrt{D}$. For a point forecast the score equals the VRMSE, so the two share one scale. Unlike a per-point CRPS, which WeatherNext 3 notes can reward members with wrong spatial covariance, it depends on the joint distribution of each field, although its sensitivity to errors in that covariance is known to be weak (Pinson & Tastu, 2013).

Protocol. Selection is on the VALID energy score with 8 fixed draws; TEST is scored once with 16 draws. The matched MLP twin shares every noise path; as in Appendix N, only the register reads the atom positions. We report:

- the energy score;
- the benchmark VRMSE of the ensemble mean;
- the spread–skill ratio (with the $(M{+}1)/M$ correction);
- the deterministic hybrid's VRMSE at the same seeds, for comparison.

The protocol was fixed before any run. Repeat runs of the deterministic hybrid at one seed differ by up to 3% on the GPU, so single-seed gaps below that are not read.

Results. Table 15 gives the vanilla-backbone rows. On every row and core the energy score is 16–20% below the deterministic hybrid's VRMSE, while the ensemble mean's VRMSE is 1–9% worse. The benchmark metric therefore does not improve; the gain is probabilistic. The spread is calibrated on shear_flow (spread–skill 1.03–1.07) and under-dispersed on TRL2D (0.57–0.58), where it can only come through the 30 routed modes.

Much of the gain comes from calibrated spread, and on TRL2D a control without training does better. A stationary Gaussian dressing of the deterministic hybrid, tuned on VALID, reaches 0.163 (0.175 on the routed modes only). The registered primary control sizes the noise on each mode by that mode's TRAIN residual relative to the window's input amplitude (the standard deviation of the last input frame), times one scale chosen on VALID. On TRL2D it reaches 0.147 with either core, below the drawn configurations even when their spread is scaled about the ensemble mean by a VALID-tuned factor τ (0.152 register, 0.158 MLP, $\tau = 2.5$ on every seed). That τ overshoots the spread–skill ratio (1.2–1.5) and still lowers the score, which suggests the drawn spread has the wrong profile across windows rather than the wrong size. On shear_flow the order reverses on every seed and core (0.017 and 0.018 drawn, 0.019 dressed), as it does on rayleigh_benard (0.035 drawn, 0.037 dressed), while on MHD_64 the dressing wins again (0.128 against 0.139–0.142). The dressing thus wins on the two rows whose remaining error the qubits cannot capture ($\eta \approx 0$, Appendix N) and loses on the two where they capture most of what the basis allows, a pattern over four rows only. Among drawn configurations the register was below its matched MLP on each of the first three seeds on TRL2D and shear_flow (paired $p = 0.055$ and 0.069). An extension to six seeds, registered after seeing those three, decides this by a two-sided paired t test with Holm correction across the two rows and requires the register lower on at least five seeds. On shear_flow it is lower on 6/6 seeds, by

3.4% (Holm $p = 0.044$), which meets the rule (0.0171, against 0.0177 for the MLP, 0.019 dressed and 0.021 deterministic). On TRL2D it is lower on 5/6 seeds by 0.8% ($p = 0.12$), which does not meet the rule, and the untrained dressing beats both cores there. The matched MLP, however, has 199 trainable parameters to the register's 244, so a second control, registered before it ran, widened the MLP by one hidden unit (297 parameters). Against it the register is lower on only 4/6 seeds on shear_flow (2.2%, $p = 0.24$), which fails the same rule: the shear advantage does not survive a wider MLP, and no core effect is claimed. The two MLP widths differ by no more than seed noise (3/6, $p = 0.48$). On TRL2D, reported without a claim, the register is lower than the wider MLP on 5/6 seeds (1.1%, $p = 0.09$).

Table 15: Distribution of configurations per input (30 qubits, TEST, mean over seeds). ES / det.: energy score over the deterministic hybrid's VRMSE, same core and seeds. Last two columns (seeds in brackets): the energy score with the spread scaled by the VALID-tuned τ, and the window-scaled Gaussian dressing of the deterministic hybrid.

Row	core	seeds	deterministic VRMSE	energy score	ES / det.	ensemble-mean VRMSE	spread–skill	ES, tuned τ	dressed (e)
TRL2D	register	3	0.1922	0.1593	0.829	0.1941	0.58	0.1517 (3)	0.1469 (3)
	MLP	3	0.1922	0.1612	0.839	0.1944	0.57	0.1578 (3)	0.1470 (3)
	register vs MLP on the energy score: +1.2% (3/3 seeds, paired $p = 0.06$); tuned τ: +3.9% (3/3, $p = 0.08$)								
shear_flow	register	3	0.0207	0.0170	0.822	0.0226	1.03	0.0171 (3)	0.0189 (3)
	MLP	3	0.0213	0.0179	0.839	0.0231	1.07	0.0179 (3)	0.0190 (3)
	register vs MLP on the energy score: +4.6% (3/3 seeds, paired $p = 0.07$); tuned τ: +4.5% (3/3, $p = 0.14$)								
rayleigh_benard†	register	3	0.0426	0.0352	0.827	0.0444	0.77	0.0350 (3)	0.0368 (3)
	MLP	3	0.0428	0.0356	0.831	0.0447	0.88	0.0358 (3)	0.0369 (3)
	register vs MLP on the energy score: +1.0% (2/3 seeds, paired $p = 0.36$); tuned τ: +2.3% (3/3, $p = 0.06$)								
MHD_64	register	3	0.1733	0.1419	0.818	0.1761	0.63	0.1419 (3)	0.1280 (3)
	MLP	3	0.1733	0.1392	0.803	0.1769	0.79	0.1387 (3)	0.1281 (3)
	register vs MLP on the energy score: -1.9% (1/3 seeds, paired $p = 0.34$); tuned τ: -2.3% (1/3, $p = 0.33$)								

P VEHICLE ROUTING: THE ROUTED REGISTER ON ROUTEFINDER

Row. RouteFinder (Berto et al., 2025) is a pretrained transformer for vehicle routing. Its released RF-TE checkpoint reproduces on the 16 variants with 100 customers to a mean gap of 3.126% (published 3.124%). The row we use is its efficient-adapter-layer (EAL) fine-tuning to 16 unseen multi-depot variants, with their protocol:

- 10 epochs of 10,000 generated instances, batch 128, learning rate 3×10^{-4}, last epoch kept;
- their test files and their gap to reference solutions;
- fp32 for every arm.

The fixed classical part ($K = 0$) is the pretrained model fine-tuned this way. A second routing anchor reproduces as well: the released RRNCO checkpoint (Son et al., 2026), which routes on real road-network distance matrices, scores 72.160, 23.279 and 73.005 on its three asymmetric CVRP test sets (in distribution, clustered, out of distribution; published 72.145, 23.282 and 73.010). Attaching the routed register to it needs its 200-epoch training schedule and is left to future work.

Routed register. A pooled insertion would see only the mean node embedding (Proposition 1). The routed register instead acts per instance:

- a learned scorer picks K nodes;
- their atoms sit at those customers' own coordinates, scaled to lattice units, plus a bounded learned offset in x, y, z;
- groups of at most 10 atoms are joined by tethers whose strengths are set per instance;
- each atom's readout ($\langle Z\rangle, \langle X\rangle, \langle Y\rangle$) is written back to its own node through a zero-initialized map, so training starts at the pretrained model.

The matched MLP uses the same routing, geometry and write-back.

Two fixes before the reported runs. First, the multi-depot environment can reach a state with no feasible action. When a remaining customer cannot be served from any depot, each depot deadlock marks the current depot unavailable until none is left, and sampling then fails. We let such rows visit

any unvisited customer. The change acts only where the original code would stop, so it leaves every other step as it was. Second, every arm trains on the same generated instances, because instance differences alone moved one pilot arm by 1.6 points.

Results. Table 16 and Figure 16 give the qubit ladder at two budgets. At the pilot budget the register is ahead of vanilla at every budget tried, and ahead of its matched MLP at 5, 10 and 30 qubits but behind it at 20. At the EAL budget no routed arm improves on the fine-tuned model alone: vanilla 8.53%, matched MLP 8.56% and register 8.59% at 30 qubits; at 20 qubits both are worse, register 8.87% and matched MLP 8.71% (seed 1). The per-variant counts in Table 16 are not independent evidence, because all 16 variants are scored by one trained policy and move together from run to run. The early lead does not survive full training, and, as on the physics rows, a matched classical core does what the register does.

Table 16: RouteFinder EAL, 16 multi-depot variants: mean test gap to the reference solutions (lower is better). Every arm trains on the same generated instances. Pilot: 1 epoch of 2,000 instances.

Budget	arm	seeds	mean gap (%)	variants better than vanilla
pilot	vanilla (K=0)	1	24.48	—
	register, K=5	1	23.66	14 of 16
	matched MLP, K=5	1	23.80	15 of 16
	register, K=10	1	24.31	11 of 16
	matched MLP, K=10	1	24.98	2 of 16
	register, K=20	1	24.36	11 of 16
	matched MLP, K=20	1	23.61	14 of 16
	register, K=30	1	23.91	13 of 16
	matched MLP, K=30	1	25.05	4 of 16
EAL protocol	vanilla (K=0)	1	8.53	—
	register, K=20	1	8.87	0 of 16
	matched MLP, K=20	1	8.71	3 of 16
	register, K=30	1	8.59	6 of 16
	matched MLP, K=30	1	8.56	8 of 16

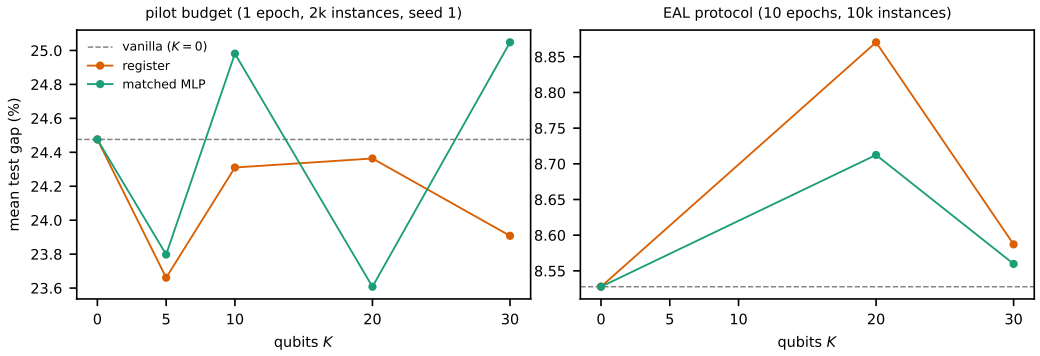

Figure 16: RouteFinder qubit ladder with the pretrained model fixed as the classical part ($K = 0$, dashed). Left: pilot budget. Right: EAL protocol. Dots are seeds.

Q Future work: foundation video world models and long-horizon generative metrics

While the paper's title and framework refer to *physics world models*, our empirical results in this work focus on continuous neural PDE surrogates (The Well, RealPDEBench) and physical transport simulators (TokaMark, TORAX), with one discrete routing row (Appendix P). We explicitly leave the evaluation of open-domain, generative video world models and their long-horizon benchmarks as future work.

Video world model benchmark suites. Emerging benchmarks evaluate whether generative video models capture intuitive physical laws from visual inputs. Specifically, Li et al. (2025) introduce *WorldModelBench*, using visual question-answering judges across physical scenarios, while Punzo et al. (2026) develop *Physics-IQ* to score video prediction models against counterfactual physical violations. Open-weights video foundation models such as Wan2.2 (Team Wan et al., 2025) and LingBot-World (Gao et al., 2026), and the 7B world-action model FLUX 3 Action (Black Forest Labs, 2026), are multi-billion-parameter diffusion or flow-matching models. FLUX 3 Action decodes future video frames and robot actions from the same future tokens; that shared latent is where the co-processor would be inserted, which we have not tested. On our hardware, one Physics-IQ pass costs 13–27 GPU-hours for three arms, and WorldModelBench needs a VLM judge pipeline on top. We make no empirical claims on video world model benchmarks in this paper; doing so under matched classical controls is an active direction for future work.

Long-horizon rollouts. The one candidate regime, rollout-aware fine-tuning evaluated on long-horizon rollouts (§5.2), did not replicate at four seeds; more seeds and a checkpoint rule that early checkpoints cannot dominate are the first requirement. Generative video world models drift away from physically consistent states over multi-second rollouts. Testing a routed register in the latent bottleneck of such a model, under the same matched controls, is future work.

Predicting the qubit scaling law. Appendix N measures the hybrid's error against the qubit budget and bounds it by a data-derived coverage curve; how efficiently qubits fill that curve is measured, not predicted. Recent theory could predict it. For an analog register the accessible frequencies are gaps of its Hamiltonian's spectrum, so per-input atom positions give input-conditioned frequency support (Wang et al., 2026b); Fourier-coefficient variance decays with frequency under non-linear embeddings such as our $\tanh$ detunings, a smooth bias consistent with our probe (Kang et al., 2026); pulse-level control changes the optimization landscape more than the reachable function class (Strobl et al., 2026); and finite-shot generalization bounds fix how a readout budget should be split between inputs and shots (Catak, 2026). Raising the coverage ceiling needs a richer or input-adaptive mode basis rather than more qubits.

Long-horizon metrics. One-step scores reward the behaviour we measure in the incoherent bands (Appendix G): where the prediction's coherence with the true change is low, the least-squares answer shrinks that band, and over a rollout the shrinkage compounds into lost small-scale energy. Long-horizon evaluation of world models should therefore report, per Fourier octave and horizon, the prediction-to-truth energy ratio and the coherence, the lag-one pole of each band's change (truth against model), and a proper probabilistic score (CRPS) for ensembles, next to each benchmark's own rollout metric. We have not run these; they are future work.