VARIATIONAL RELEVANCE-INFORMED SPARSIFICATION OF TERM MEASUREMENTS FOR VARIATIONAL QUANTUM ALGORITHMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated measurements are a major bottleneck in variational quantum algorithms (VQAs). We uncover and characterize parameter-specific sparsity in the matrix of qubit-wise commuting (QWC) group contributions to parameter-shift gradients: most absolute contribution mass concentrates in a subset of groups whose relevance varies across parameters and optimization states. Building on this structure, we introduce a learned controller that predicts group relevance from Hamiltonian, circuit, light-cone, and current-state features, then selects measurements through a tunable coverage threshold. Exact contributions provide offline supervision without being required at deployment. Across 15 molecular VQE settings, our method saves 49.6% of nominal gradient-measurement shots with a mean absolute energy gap of 5.67 mHa to Full-QWC. We then ask whether large language models (LLMs) can solve the same selection problem without fine-tuning. The results reveal both promise and a clear limitation: LLMs can identify useful measurement groups, yet without guidance, local models do not consistently prioritize the physical information that matters most. Our learned controller uses these cues more effectively. This contrast points toward a broader direction for AI-assisted science: combining the flexibility of general-purpose models with explicit physical representations, task-specific objectives, and domain-grounded validation.

1 INTRODUCTION

The variational quantum eigensolver (VQE) is the flagship variational quantum algorithm (VQA) for estimating molecular ground-state energies (Peruzzo et al., 2014; McClean et al., 2016). Much of the research on VQAs has focused on trainability, particularly on mitigating barren plateaus (BPs) (McClean et al., 2018). Measurement overhead presents an equally fundamental challenge, yet has received comparatively less attention. Even with hardware-efficient circuits, each VQE iteration requires repeated estimation of Hamiltonian observables and gradients. For quantum chemistry, this measurement burden often dominates the practical cost (McArdle et al., 2020; Gonthier et al., 2022).

To identify where this measurement burden can be reduced, we first decompose the cost of a single gradient update. Because unknown quantum states cannot be cloned and quantum measurements are destructive (Wootters & Zurek, 1982), expectation values must be estimated through repeated circuit preparations. On this sampling-based interface, parameter-shift rules are widely used because they provide hardware-compatible analytic gradients (Schuld et al., 2019). Let a qubit Hamiltonian contain G QWC measurement groups and the ansatz P trainable parameters. A parameter-shift gradient evaluates two shifted circuits per parameter and, under the standard Full-QWC estimator, measures all G groups with S shots, so one optimization step costs $2PGS$ circuit shots. Existing methods shrink one factor at a time: grouping compatible Pauli operators reduces G (Gokhale et al., 2019; Crawford et al., 2021), adaptive shot allocation reduces S (Kübler et al., 2020; Gu et al., 2021; Zhu et al., 2024), operator sampling subsamples the Hamiltonian (Arrasmith et al., 2020), and parameter freezing reduces P (Peng et al., 2026). None of these asks the finer question: *when parameter j is evaluated, which QWC groups are actually needed for its gradient?*

The answer is strongly parameter dependent: every parameter sees the same Hamiltonian groups, yet the shifted expectation differences depend on gate location, ansatz light cone, and current circuit

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.

state, so the contribution matrix over parameter–group pairs is concentrated but not captured by any global coefficient ranking. Exploiting this structure faces a circularity: computing the matrix exactly at each iteration would require precisely the Full-QWC measurements we seek to avoid.

Our method resolves the circularity by learning to predict the latent matrix. Separate Transformer encoders contextualize parameter and QWC-group tokens; a pairwise scorer estimates a normalized relevance distribution for every parameter; a row-wise top-mass rule converts the prediction into a measurement mask. τ thereby becomes a direct post-training accuracy-cost control; exact contributions are used only offline, never at deployment. Our contributions are threefold:

- We uncover parameter-specific sparsity in the QWC contribution matrix of variational quantum gradients in Figure 1: most absolute contribution mass concentrates in a subset of measurement groups whose relevance varies across parameters and optimization states.
- We develop a size-flexible learned controller that directly predicts parameter–group relevance from physical and current-state features, using exact contributions only as offline supervision. A single threshold converts these predictions into measurement subsets, enabling adjustable accuracy–cost trade-offs without retraining.
- We investigate whether LLMs can perform this scientific selection task without fine-tuning. Frontier models achieve millihartree-scale gaps with retrieved training examples, while local-model audits reveal uneven utility of the supplied physical information. These findings demonstrate the potential of LLMs for scientific decision making and motivate domain-specific guidance as a promising route toward more reliable AI-assisted research.

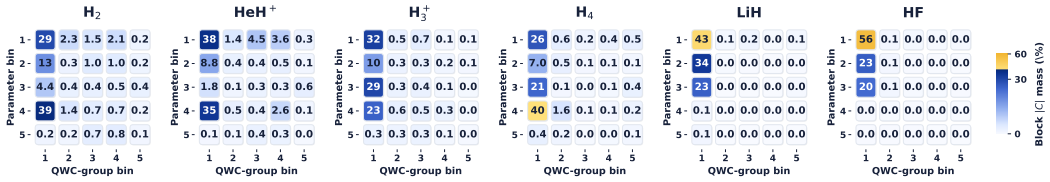

Figure 1: **Parameter-specific QWC sparsity.** Cells show contribution-mass percentages in a 5×5 visualization of the exact parameter–group matrix; training and evaluation use the unaggregated matrix. The corresponding result under an IBM `FakeGuadalupeV2` calibration-derived noise model (IBM Quantum, 2026a;b) is reported in Appendix K, Figure 23.

Dominant groups vary by parameter, yet each row concentrates on a small fraction of the measurement set; this is the sparse structure our method exploits, and it persists under the calibrated-noise audit in Appendix K, indicating that it is not merely an artifact of idealized, noiseless execution.

2 Problem Formulation and Method

A molecular electronic-structure Hamiltonian mapped to qubits can be written as

$$H = h_0 I + \sum_{\ell=1}^{L} h_\ell P_\ell, \qquad E(\boldsymbol{\theta}) = \langle \psi(\boldsymbol{\theta}) | H | \psi(\boldsymbol{\theta}) \rangle, \tag{1}$$

where P_ℓ is a Pauli string and h_ℓ is fixed. Qubit-wise commuting strings share a tensor-product basis and form groups $\{\mathcal{G}_k\}_{k=1}^{G}$ (Verteletskyi et al., 2020; Crawford et al., 2021).

For RY/RZ gates, the parameter-shift rule gives (Schuld et al., 2019; Wierichs et al., 2022)

$$g_j(\boldsymbol{\theta}) = \frac{\partial E}{\partial \theta_j} = \frac{1}{2}\left[E\Big(\boldsymbol{\theta} + \frac{\pi}{2}\boldsymbol{e}_j\Big) - E\Big(\boldsymbol{\theta} - \frac{\pi}{2}\boldsymbol{e}_j\Big)\right]. \tag{2}$$

Expanding each shifted energy over QWC groups exposes the signed contribution of group k to j:

$$C_{j,k}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{\ell \in \mathcal{G}_k} h_\ell \left[\langle P_\ell \rangle_{\boldsymbol{\theta} + \frac{\pi}{2}\boldsymbol{e}_j} - \langle P_\ell \rangle_{\boldsymbol{\theta} - \frac{\pi}{2}\boldsymbol{e}_j}\right], \qquad g_j = \sum_{k=1}^{G} C_{j,k}. \tag{3}$$

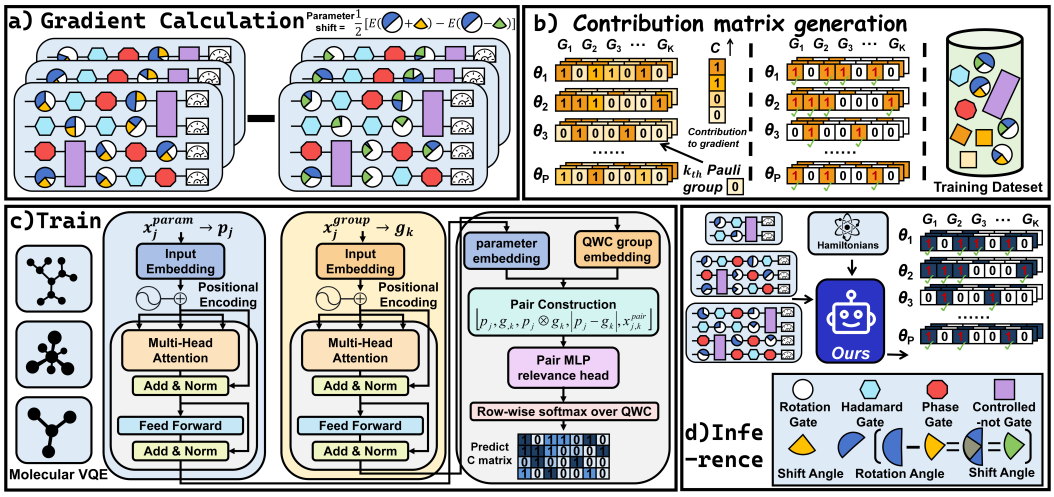

Figure 2: **Method overview in four stages.** (a) A Hamiltonian is partitioned into QWC groups and evaluated with a parameterized ansatz. (b) Full-QWC parameter shift measures every group for every parameter at cost $2PGS$. (c) The learned controller predicts the parameter–group relevance matrix and retains the smallest per-parameter subset covering top-τ mass. (d) Closed-loop VQE measures only those groups, forms sparse gradients, and updates all parameters at reduced cost.

Figure 2 summarizes our framework containing 4 stages. The coefficients and QWC partition stay fixed, but $C_{j,k}$ changes with the ansatz state. At an offline training snapshot, we evaluate Eq. 3 and normalize its absolute contribution,

$$w_{j,k}(\boldsymbol{\theta}) = \frac{|C_{j,k}(\boldsymbol{\theta})|}{\sum_{r=1}^{G} |C_{j,r}(\boldsymbol{\theta})|}. \tag{4}$$

Exact C supplies offline labels and simulator diagnostics; selectors never receive current oracle contributions. Separate Transformer encoders (Vaswani et al., 2017) contextualize parameter and group tokens (position, angle/time, coefficient, Pauli support, light cones). A pair scorer receives

$$z_{j,k} = [p_j, q_k, p_j \odot q_k, |p_j - q_k|, x_{j,k}^{\text{pair}}], \tag{5}$$

and predicts row-normalized relevance $\widehat{w}_{j,k}$. Training combines distribution, top-set, ranking, budget, and entropy losses:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{m}}\mathcal{L}_{\text{mask}} + \lambda_{\text{r}}\mathcal{L}_{\text{rank}} + \lambda_{\text{b}}\mathcal{L}_{\text{budget}} + \lambda_{\text{e}}\mathcal{L}_{\text{entropy}}. \tag{6}$$

For each parameter, our method chooses the smallest set $S_j(\tau)$ satisfying $\sum_{k\in S_j(\tau)} \widehat{w}_{j,k} \geq \tau$.

Only groups in $S_j(\tau)$ are measured on the two shifted circuits; with S shots per setting,

$$B_{\text{Full}} = 2SPG, \qquad B_{\text{Ours}} = 2S\sum_{j=1}^{P} |S_j(\tau)|, \qquad \text{gain} = 1 - \frac{B_{\text{Ours}}}{B_{\text{Full}}}. \tag{7}$$

For fixed predictions, increasing τ nests the selected masks and directly controls measurement coverage. In closed-loop optimization, each mask also shapes subsequent circuit states. The omitted signed contributions determine the truncation bias; Appendix M gives its bound, the complete losses, and implementation details. Appendix N.2 identifies the checkpoint used in each sweep. The formulation covers Pauli-observable VQAs; related work is in Appendix A.

3 Experimental Setup

PySCF/OpenFermion STO-3G Hamiltonians (Sun et al., 2018; McClean et al., 2020; Hehre et al., 1969) for H2, HeH+, H3+, H4, and LiH map under Jordan–Wigner (Seeley et al., 2012) to 4–12 qubits and 5–179 QWC groups; the main ansatz is layered RY/RZ with linear entanglers (Kandala et al., 2017), and structural-OOD tests use named ansatze and spin Hamiltonians.

Table 1: **Cost-matched online VQE at $\tau = 0.995$.** Exact-expectation results; shot gains are nominal. Entries are mean $\pm$ standard deviation over four seeds; ΔE is in Ha. Random and coefficient controls use comparable budgets; TITAN† is cell-wise shot matched. Green/red shading indicates the one-sided 1.6 mHa criterion. **Bold** marks the lowest $|\Delta E|$ or largest gain; underlining marks the second-largest gain. Metrics, full controls, and noise extensions are in Appendices M.4, B, and K.

Setting	Metric	Method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 1	ΔE vs Full ↓	**Ours**	**+0.0011 ± 0.0008**	**+0.0015 ± 0.0012**	**+0.0005 ± 0.0006**	**−0.0006 ± 0.0024**	+0.0023 ± 0.0016
Depth 1	ΔE vs Full ↓	Random	+0.2359 ± 0.0507	+1.2220 ± 0.2247	+0.8916 ± 0.1887	+1.4117 ± 0.2745	+3.4982 ± 0.0793
Depth 1	ΔE vs Full ↓	Global coeff.	+0.0038 ± 0.0053	+0.0019 ± 0.0041	+0.0122 ± 0.0093	+0.0089 ± 0.0074	**+0.0016 ± 0.0026**
Depth 1	ΔE vs Full ↓	TITAN†	+0.0128 ± 0.0170	+0.1299 ± 0.0870	+0.0772 ± 0.0483	−0.0642 ± 0.0758	−0.0648 ± 0.1287
Depth 1	ΔE vs Full ↓	TITAN†**+Ours**	+0.0127 ± 0.0160	+0.1297 ± 0.0845	+0.0776 ± 0.0483	−0.0649 ± 0.0752	−0.0646 ± 0.1286
Depth 1	Shot gain ↑	**Ours**	40.4 ± 6.3%	58.9 ± 3.5%	63.0 ± 2.8%	74.7 ± 1.1%	86.9 ± 0.5%
Depth 1	Shot gain ↑	Random	42.0 ± 4.6%	60.7 ± 1.5%	66.5 ± 2.2%	79.2 ± 1.3%	90.2 ± 0.3%
Depth 1	Shot gain ↑	Global coeff.	41.7 ± 6.5%	60.0 ± 3.3%	62.9 ± 2.9%	75.0 ± 0.3%	87.0 ± 0.6%
Depth 1	Shot gain ↑	TITAN†	43.8 ± 6.2%	59.4 ± 6.3%	62.7 ± 6.4%	70.3 ± 5.7%	81.8 ± 5.2%
Depth 1	Shot gain ↑	TITAN†**+Ours**	**66.8 ± 3.7%**	**91.7 ± 2.0%**	**95.0 ± 2.7%**	**97.9 ± 0.9%**	**99.9 ± 0.1%**
Depth 2	ΔE vs Full ↓	**Ours**	**+0.0003 ± 0.0007**	−0.0011 ± 0.0026	**−0.0015 ± 0.0041**	**+0.0158 ± 0.0443**	**−0.0049 ± 0.0097**
Depth 2	ΔE vs Full ↓	Random	+0.1656 ± 0.1105	+0.4704 ± 0.1737	+0.6527 ± 0.0767	+0.9343 ± 0.1414	+4.0043 ± 0.3648
Depth 2	ΔE vs Full ↓	Global coeff.	+0.0015 ± 0.0091	**−0.0004 ± 0.0088**	+0.0233 ± 0.0429	+0.0379 ± 0.0540	−0.0049 ± 0.0123
Depth 2	ΔE vs Full ↓	TITAN†	+0.0075 ± 0.0093	+0.0116 ± 0.0467	+0.0153 ± 0.0618	+0.0472 ± 0.0558	+0.2903 ± 0.1308
Depth 2	ΔE vs Full ↓	TITAN†**+Ours**	+0.0072 ± 0.0082	+0.0129 ± 0.0483	+0.0140 ± 0.0616	+0.0493 ± 0.0560	+0.2920 ± 0.1356
Depth 2	Shot gain ↑	**Ours**	25.6 ± 2.9%	31.6 ± 2.6%	44.9 ± 1.3%	51.8 ± 0.5%	82.6 ± 0.4%
Depth 2	Shot gain ↑	Random	26.3 ± 2.5%	33.2 ± 1.5%	47.7 ± 2.1%	53.7 ± 0.4%	83.4 ± 1.3%
Depth 2	Shot gain ↑	Global coeff.	25.7 ± 2.8%	31.9 ± 2.4%	44.7 ± 0.8%	51.7 ± 0.5%	82.5 ± 0.3%
Depth 2	Shot gain ↑	TITAN†	23.8 ± 4.7%	39.0 ± 14.5%	41.9 ± 9.4%	46.8 ± 9.9%	82.7 ± 5.3%
Depth 2	Shot gain ↑	TITAN†**+Ours**	**40.2 ± 5.9%**	**64.0 ± 10.6%**	**74.5 ± 6.6%**	**79.0 ± 6.3%**	**99.3 ± 0.3%**

Training uses depths 1-3 and 8 seeds. Four held-out seeds run 28 Adam steps with exact masked expectations and nominal cost at 1024 shots/setting. We report $\Delta E = E_{\text{method}} - E_{\text{Full}}$ (Ha) and shot gain; *no worse* means $\Delta E \leq 1.6$ mHa. Metric and estimator definitions are in Appendix M.4; 400-step shot-noise checks are in Appendix L.

4 Main Results

At $\tau = 0.995$, our method saves 49.6% of parameter-shift shots at a 5.67 mHa seed-level mean absolute gap, attains the smallest $|\Delta E|$ in eight of ten molecule–depth cells of Table 1, and satisfies the one-sided no-worse criterion in 8/10 cells, more than any alternative (Appendix B, Table 5). Tightening to $\tau = 0.997$ reduces the mean gap to 3.52 mHa while still saving 45.5%; the complete τ frontier, including its behavior under depth shift, is in Appendix C (Figure 8).

Each control isolates one candidate explanation of the savings. Count-matched Random inherits our method's per-parameter budgets and attains slightly *larger* shot gains in every cell, yet drifts by 0.17-4.00 Ha: the savings come from the budget size, while the accuracy comes from the learned ranking. Static coefficient ranking (natural training-free ordering) performs no worse in only 3/10 cells. Shot-matched TITAN† freezing (Peng et al., 2026) drifts by tens to hundreds of mHa, so the gain is not explained by evaluating fewer parameters; conversely, TITAN†**+Ours** achieves the largest gain, showing that relevance selection composes with, but does not repair, aggressive freezing.

The endpoint and trajectory views resolve complementary aspects of the trade-off. LiH depth 1 lies 0.7 mHa above the no-worse threshold, while H4 depth 2 ($+15.8 \pm 44.3$ mHa) exhibits strong seed dependence. Exact-expectation trajectories (Figure 7) reach Full-QWC parity at 74.4% (H2) and 55.1% (H3+) of its nominal budget. Separate 400-step runs with simulated shot noise (Appendix L) keep every molecule's mean final gap within 0.18 mHa of Full-QWC while saving 24.7–92.1%.

The learned ranking is also not tied to one architecture: swapping the Transformer for message-passing GNN (Gilmer et al., 2017), MLP, or ResNet-18 (He et al., 2016) predictors (Appendix I) preserves the qualitative picture, while the Transformer has the lowest tail error on the measurement-intensive molecules and the strongest molecule-OOD transfer. Architecture details are in Appendix M.3. The consistency across model families supports parameter-group relevance as a useful supervised learning target. The Transformer's lower tail error further motivates its use as our default predictor, with robustness across molecular systems guiding the architectural choice.

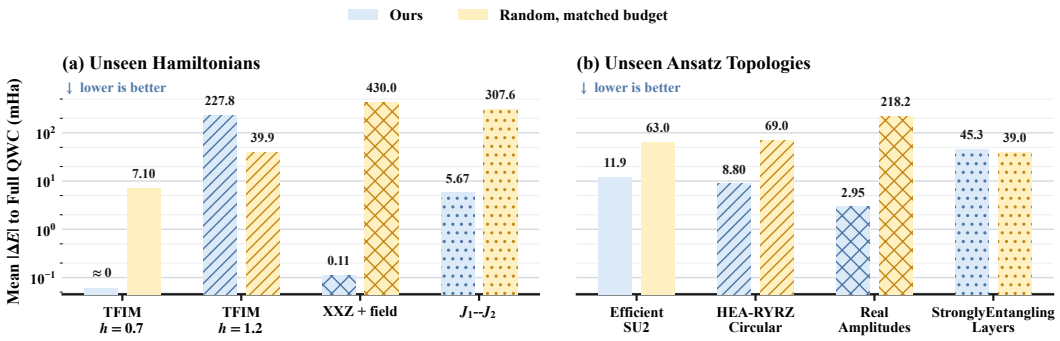

Figure 3: **Matched-budget structural OOD.** Mean absolute gaps for unseen Hamiltonian families (a) and ansatz topologies (b); Random receives the same parameter-specific group counts as our method. Exact values and gains are in Appendix F.

5 Allocation and Structural Generalization

Table 2: **Shot-allocation Pareto benchmark.** Five molecules, depths 1–2, and four seeds under common Adam with simulated shot noise. ΔE is in Ha; gain is relative to uniform Full-QWC at budget b; no-worse means $\Delta E \leq 1.6$ mHa. **Bold**/underline mark best/second-best non-Full values. Adapted-policy settings and detailed results are in Appendices N.1 and D.

Method	$b = 0.25$ Mean $\lvert\Delta E\rvert \downarrow$	Shot gain ↑	No-worse ↑	$b = 0.50$ Mean $\lvert\Delta E\rvert \downarrow$	Shot gain ↑	No-worse ↑	$b = 1.00$ Mean $\lvert\Delta E\rvert \downarrow$	Shot gain ↑	No-worse ↑
Full-QWC uniform	0.00000	0.0%	100.0%	0.00000	0.0%	100.0%	0.00000	0.0%	100.0%
iCANS (Kübler et al., 2020)	0.17283	−35.0%	10.0%	0.11977	−49.8%	37.5%	0.09657	−76.6%	17.5%
gCANS (Gu et al., 2021)	0.20449	92.9%	7.5%	0.14059	92.2%	25.0%	0.10789	91.8%	27.5%
Rosalin (Arrasmith et al., 2020)	0.06008	−78.6%	45.0%	0.09137	−90.9%	72.5%	0.06912	−97.8%	67.5%
Variance-aware (Zhu et al., 2024)	0.06694	0.0%	65.0%	0.06521	0.0%	**80.0%**	0.06278	0.0%	72.5%
Ours + uniform	**0.05472**	57.5%	**72.5%**	**0.06064**	57.1%	65.0%	0.06376	57.1%	65.0%
Ours + variance-aware	0.07211	57.1%	65.0%	0.06876	57.1%	**80.0%**	**0.06031**	56.8%	**82.5%**
Ours + gCANS	0.12300	**97.3%**	17.5%	0.16134	**97.2%**	27.5%	0.07148	**97.2%**	30.0%

Separate transfer studies test four shift axes: Figure 3 shows our method beats count-matched Random on three of four unseen Hamiltonian families and on 44/45 transferable named-ansatz cells. TFIM at h=1.2 and StronglyEntanglingLayers identify challenging structural shifts (Appendix H). Table 2 demonstrates composition with allocation policies: our method+variance-aware saves 56.8% and raises the no-worse rate from 72.5% to 82.5%. The common Adam protocol controls optimizer dynamics while comparing allocation and selection; exact policy settings are in Appendix N.1. Complete results are in Appendices F, E–H, D, and K; predictor diagnostics are in Appendix I.

6 LLM Controls and Mechanism Audit

Do the gains require task-specific learning, or can a capable general model be prompted into the same decisions? Figure 4 shows the shared interface: our method supplies each parameter's cardinality k_j and structured physical features, a frozen zero-shot LLM selects the group identities, and the validated mask is reused for K=7 steps. Matching refresh cadence controls invocation frequency. We evaluate four frozen local LLMs: Qwen3-8B (Yang et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025), and GPT-OSS-20B (Agarwal et al., 2025). In Table 3, relative to per-step **Ours**-K=1, the four-refresh **Ours**-K=7 control increases the cross-cell mean absolute gap from 0.00567 to 0.00715 Ha while changing mean shot saving from 49.6% to 51.4%; stale masks do cost reliability (the no-worse count falls

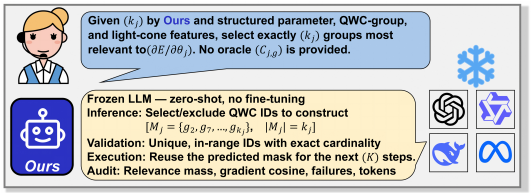

Figure 4: **Clean frozen-LLM protocol.** Our learned controller supplies k_j and features; the LLM returns select/exclude QWC IDs; the mask is reused for K=7.

Table 3: **Frozen local-LLM QWC ranking.** Four zero-shot LLMs receive predicted counts and refresh every $K=7$ steps. Entries are mean ± s.d. over four seeds (ΔE: Ha; gain: %). Green/red shading indicates the 1.6 mHa no-worse split; ✓/✗ compares $|\Delta E|$ with Random. **Bold**/underline mark the smallest/second-smallest $|\Delta E|$ per cell. Protocols are in Appendices J and N.4.

Setting	Ranking method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 1	**Ours** gain, $K=1$ ↑	40.4 ± 6.3%	58.9 ± 3.5%	63.0 ± 2.8%	74.7 ± 1.1%	86.9 ± 0.5%
Depth 1	**Ours** gain, $K=7$ ↑	41.2 ± 5.3%	60.7 ± 3.1%	65.1 ± 2.6%	77.3 ± 0.7%	88.2 ± 0.4%
Depth 1	Random ($K=7$)	+0.1847 ± 0.0868	+1.2191 ± 0.2759	+0.7247 ± 0.2856	+0.9736 ± 0.3101	+2.9030 ± 0.6058
Depth 1	Qwen3-8B	+0.1321 ± 0.0938 ✓	+1.3928 ± 0.9006 ✗	+0.9972 ± 0.1834 ✗	+0.4765 ± 0.2133 ✓	+2.0656 ± 0.7546 ✓
Depth 1	Llama-3.1-8B	+0.2840 ± 0.1350 ✗	+1.2347 ± 0.3492 ✗	+0.6195 ± 0.3195 ✓	+0.4219 ± 0.3064 ✓	+1.3266 ± 0.8740 ✓
Depth 1	DeepSeek-R1-8B	+0.2405 ± 0.1463 ✗	+1.0993 ± 0.5373 ✓	+0.2766 ± 0.1157 ✓	+0.0522 ± 0.0653 ✓	+0.0504 ± 0.0576 ✓
Depth 1	GPT-OSS-20B	+0.0471 ± 0.0463 ✓	+0.2783 ± 0.2758 ✓	+0.0568 ± 0.0867 ✓	+0.0379 ± 0.0641 ✓	**+0.0018 ± 0.0032** ✓
Depth 1	**Ours** ($K=7$)	+0.0013 ± 0.0003 ✓	+0.0035 ± 0.0015 ✓	**+0.0004 ± 0.0006** ✓	**+0.0001 ± 0.0060** ✓	+0.0030 ± 0.0016 ✓
Depth 1	**Ours** ($K=1$)	**+0.0011 ± 0.0008** ✓	**+0.0015 ± 0.0012** ✓	+0.0005 ± 0.0006 ✓	−0.0006 ± 0.0024 ✓	+0.0023 ± 0.0016 ✓
Depth 2	**Ours** gain, $K=1$ ↑	25.6 ± 2.9%	31.6 ± 2.6%	44.9 ± 1.3%	51.8 ± 0.5%	82.6 ± 0.4%
Depth 2	**Ours** gain, $K=7$ ↑	27.3 ± 2.5%	32.5 ± 2.4%	47.9 ± 1.4%	55.1 ± 0.5%	83.7 ± 0.4%
Depth 2	Random ($K=7$)	+0.0930 ± 0.0297	+0.5668 ± 0.4463	+0.3054 ± 0.1633	+0.9327 ± 0.3273	+2.6969 ± 0.2816
Depth 2	Qwen3-8B	+0.0992 ± 0.0689 ✗	+0.8415 ± 0.1479 ✗	+0.6751 ± 0.2741 ✗	+0.1842 ± 0.0985 ✓	+2.6722 ± 0.9356 ✓
Depth 2	Llama-3.1-8B	−0.0595 ± 0.1790 ✓	+0.2644 ± 0.2592 ✓	+0.6780 ± 0.0990 ✗	−0.0543 ± 0.0502 ✓	+0.4541 ± 0.2847 ✓
Depth 2	DeepSeek-R1-8B	+0.0051 ± 0.1945 ✓	+0.0991 ± 0.0653 ✓	+0.3239 ± 0.1195 ✗	+0.0167 ± 0.0328 ✓	+0.0332 ± 0.0342 ✓
Depth 2	GPT-OSS-20B	+0.0011 ± 0.0098 ✓	+0.0049 ± 0.0044 ✓	+0.0322 ± 0.0484 ✓	+0.0167 ± 0.0328 ✓	+0.0603 ± 0.1773 ✓
Depth 2	**Ours** ($K=7$)	−0.0007 ± 0.0015 ✓	+0.0076 ± 0.0176 ✓	**+0.0006 ± 0.0031** ✓	+0.0162 ± 0.0438 ✓	**−0.0040 ± 0.0118** ✓
Depth 2	**Ours** ($K=1$)	**+0.0003 ± 0.0007** ✓	**−0.0011 ± 0.0026** ✓	−0.0015 ± 0.0041 ✓	**+0.0158 ± 0.0443** ✓	−0.0049 ± 0.0097 ✓
Depth 3	**Ours** gain, $K=1$ ↑	18.1 ± 2.7%	24.9 ± 1.3%	29.9 ± 1.0%	35.2 ± 0.9%	75.4 ± 1.8%
Depth 3	**Ours** gain, $K=7$ ↑	19.0 ± 2.4%	26.1 ± 1.2%	32.1 ± 1.2%	38.6 ± 0.7%	76.6 ± 1.3%
Depth 3	Random ($K=7$)	+0.0730 ± 0.0859	+0.2617 ± 0.2130	+0.2750 ± 0.0765	+0.2773 ± 0.2665	+2.3928 ± 0.4840
Depth 3	Qwen3-8B	+0.0742 ± 0.0922 ✗	+0.8192 ± 0.3035 ✗	+0.3073 ± 0.1583 ✗	+0.0046 ± 0.1338 ✓	+0.9203 ± 0.3562 ✓
Depth 3	Llama-3.1-8B	+0.0631 ± 0.0983 ✓	+0.3042 ± 0.1041 ✗	+0.1583 ± 0.0876 ✓	−0.0230 ± 0.0343 ✓	−0.0493 ± 0.3344 ✓
Depth 3	DeepSeek-R1-8B	−0.0279 ± 0.0752 ✓	+0.0302 ± 0.0339 ✓	+0.0823 ± 0.0558 ✓	−0.0093 ± 0.0205 ✓	+0.0306 ± 0.0379 ✓
Depth 3	GPT-OSS-20B	+0.0436 ± 0.0577 ✓	+0.1632 ± 0.1611 ✓	+0.0091 ± 0.0196 ✓	−0.0033 ± 0.0142 ✓	−0.0392 ± 0.1522 ✓
Depth 3	**Ours** ($K=7$)	+0.0010 ± 0.0027 ✓	+0.0108 ± 0.0118 ✓	+0.0029 ± 0.0095 ✓	+0.0029 ± 0.0113 ✓	−0.0046 ± 0.0040 ✓
Depth 3	**Ours** ($K=1$)	**+0.0005 ± 0.0015** ✓	**+0.0092 ± 0.0131** ✓	**+0.0005 ± 0.0077** ✓	**−0.0014 ± 0.0097** ✓	**−0.0017 ± 0.0036** ✓

from 12/15 to 8/15 cells). Even so, **Ours**-$K=7$ remains closer to Full-QWC than Qwen, Llama, and DeepSeek in all 15 cells, and closer than GPT-OSS in 13/15, at nearly identical aggregate savings. GPT-OSS is the strongest frozen LLM (0.0694 Ha mean absolute gap), followed by DeepSeek (0.172), Llama (0.421), and Qwen (0.786). Against the protocol-matched Random control, Qwen, Llama, DeepSeek, and GPT-OSS win 7/15, 11/15, 13/15, and 15/15 cells, respectively, while both proposed-method controls win 15/15: physics-conditioned prompting beats random choice, yet the aggregate LLM gap stays almost an order of magnitude above that of our refresh-matched method; the advantage comes from learned parameter-specific ranking, not from invocation frequency.

The repair-free audit (Appendix J, Table 50) jointly changes the output contract and replaces coefficient completion with Full-QWC fallback after failed retries. Models retain 52.0–52.4% mean shot saving and beat Random in 13/15, 10/15, 11/15, and 7/15 cells (Qwen, Llama, DeepSeek, GPT-OSS), with mean absolute gaps of 0.4482–0.6986 Ha. Both proposed-method controls win 15/15. This second interface tests the same scientific selection task under a distinct validation policy. Protocol versions and direct-JSON inference settings are detailed in Appendix N.4; Figure 12 retains the earlier schematic.

Data-exposed frontier models. We next test whether scale plus data access closes the gap: five GPT and four Claude variants receive two train-only same-molecule/depth/parameter neighbors and an eight-neighbor summary (retrieval details in Appendix N.3). They choose both group IDs and counts, yielding model-specific accuracy–cost operating points. The held-out contribution matrix is never shown, tools are disabled, and an invalid response is charged as Full-QWC; all methods share the 28-step, $K=7$ protocol and held-out seeds. Table 4 shows depth 1 with five GPT variants (OpenAI, 2026) and four Claude variants (Anthropic, 2026); Appendix J, Tables 51–52, give depths 2–3. Data-exposed frontier models do reach the mHa regime; the relevance structure is learnable in con-

Table 4: **Data-exposed frontier models at depth 1.** Train-only retrieval informs group IDs and counts under four-refresh $K=7$. Entries are mean ± s.d. over four seeds. Green/red shading indicates the 1.6 mHa no-worse split; ✓/✗ compares $|\Delta E|$ with Random. Among non-Random methods, **bold**/underline mark best/second-best $|\Delta E|$ ↓ and gain ↑; gray marks largest gain. Protocols and full results: Appendices N.3 and J.

Method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Energy fidelity ↓: ΔE vs. Full (Ha)					
Random ($K=7$)	+0.1847 ± 0.0868	+1.2191 ± 0.2759	+0.7247 ± 0.2856	+0.9736 ± 0.3101	+2.9030 ± 0.6058
GPT-5.4	+0.0002 ± 0.0003 ✓	−0.0007 ± 0.0017 ✓	+0.0010 ± 0.0011 ✓	+0.0021 ± 0.0015 ✓	**+0.0010 ± 0.0005** ✓
GPT-5.5	+0.0001 ± 0.0007 ✓	+0.0006 ± 0.0011 ✓	+0.0008 ± 0.0009 ✓	+0.0026 ± 0.0022 ✓	+0.0017 ± 0.0012 ✓
GPT-5.6 Sol	+0.0002 ± 0.0005 ✓	−0.0007 ± 0.0023 ✓	+0.0015 ± 0.0015 ✓	+0.0002 ± 0.0023 ✓	+0.0018 ± 0.0007 ✓
GPT-5.6 Terra	**−0.0000 ± 0.0007** ✓	+0.0009 ± 0.0015 ✓	+0.0041 ± 0.0050 ✓	+0.0017 ± 0.0033 ✓	+0.0016 ± 0.0005 ✓
GPT-5.6 Luna	+0.0003 ± 0.0004 ✓	+0.0001 ± 0.0010 ✓	+0.0011 ± 0.0013 ✓	+0.0035 ± 0.0024 ✓	+0.0018 ± 0.0009 ✓
Claude Haiku 4.5	−0.0001 ± 0.0004 ✓	+0.0002 ± 0.0004 ✓	+0.0006 ± 0.0010 ✓	+0.0020 ± 0.0033 ✓	+0.0011 ± 0.0007 ✓
Claude Sonnet 5	+0.0000 ± 0.0004 ✓	+0.0004 ± 0.0010 ✓	**+0.0003 ± 0.0010** ✓	+0.0017 ± 0.0022 ✓	+0.0016 ± 0.0005 ✓
Claude Opus 5	+0.0002 ± 0.0004 ✓	+0.0005 ± 0.0009 ✓	+0.0004 ± 0.0009 ✓	+0.0024 ± 0.0040 ✓	+0.0015 ± 0.0008 ✓
Claude Fable 5	+0.0003 ± 0.0007 ✓	**−0.0000 ± 0.0006** ✓	+0.0004 ± 0.0010 ✓	+0.0010 ± 0.0044 ✓	+0.0011 ± 0.0005 ✓
Ours ($K=7$)	+0.0013 ± 0.0003 ✓	+0.0035 ± 0.0015 ✓	+0.0004 ± 0.0006 ✓	**+0.0001 ± 0.0060** ✓	+0.0030 ± 0.0016 ✓
Measurement efficiency: shot gain (%; higher is better ↑)					
Random ($K=7$)	42.7 ± 3.8%	61.3 ± 2.6%	68.2 ± 2.5%	80.2 ± 1.2%	90.8 ± 0.1%
GPT-5.4	17.2 ± 1.0%	46.7 ± 4.2%	52.2 ± 3.7%	64.6 ± 1.9%	80.4 ± 0.3%
GPT-5.5	18.6 ± 1.6%	45.4 ± 3.8%	51.8 ± 3.3%	70.4 ± 2.6%	82.9 ± 2.2%
GPT-5.6 Sol	18.1 ± 1.4%	47.3 ± 4.5%	53.3 ± 3.7%	64.3 ± 1.2%	80.7 ± 0.3%
GPT-5.6 Terra	18.1 ± 1.0%	46.7 ± 2.3%	53.6 ± 2.5%	64.1 ± 1.1%	80.2 ± 0.4%
GPT-5.6 Luna	19.4 ± 2.5%	46.8 ± 4.1%	52.3 ± 3.6%	64.7 ± 1.0%	80.8 ± 0.3%
Claude Haiku 4.5	18.1 ± 1.6%	46.9 ± 3.3%	52.0 ± 3.1%	65.1 ± 1.8%	81.3 ± 1.1%
Claude Sonnet 5	19.7 ± 1.4%	43.8 ± 3.3%	50.3 ± 3.1%	69.6 ± 0.9%	85.7 ± 0.6%
Claude Opus 5	17.2 ± 1.1%	42.3 ± 3.1%	44.7 ± 3.0%	68.9 ± 2.2%	81.9 ± 1.0%
Claude Fable 5	18.3 ± 1.5%	44.3 ± 3.4%	50.4 ± 3.0%	67.2 ± 2.1%	80.1 ± 0.5%
Ours ($K=7$)	**41.2 ± 5.3%**	**60.7 ± 3.1%**	**65.1 ± 2.6%**	**77.3 ± 0.7%**	**88.2 ± 0.4%**

text, but the outcome is a cost–accuracy frontier rather than uniform dominance: Claude Fable 5 attains the best overall mean $|\Delta E|$ (0.00302 Ha) at 35.8% saving, **Ours**-$K=7$ operates at a more aggressive 51.4% saving with 0.00715 Ha, and no model dominates every cell.

Appendix Table 53 records 17.0–32.4M tokens per frontier model; our controller has 984,454 parameters and a 3.78-MiB checkpoint. These are distinct resource measures; shot gain excludes offline training and classical costs (Appendix N.1).

Prompt, context, and feature audit. *Which parts of the interface improve scientific selection?* We freeze 12 Full-QWC snapshots and test the four local LLMs under 15 prompt, context, output, feature-deletion, and shuffle conditions (720 cases). The oracle matrix is audit-only, and cardinality is fixed. Figure 5 shows one observable Llama decision; the corresponding Qwen, DeepSeek, and GPT-OSS traces are in Appendix Figures 13, 14, and 15. These are observable outputs paired with external feature audits. The harness performs validation and VQE updates; frozen LLMs produce structured selection JSON through text inference with tool execution disabled.

The Qwen, DeepSeek, and GPT-OSS traces (Appendix Figures 13, 14, 15) capture 95.0%, 94.4%, and 93.0% mass, versus Llama's 1.8%. Together, these cases reveal substantial model-specific differences within the same physical task. Appendix N.4 links each action to its prompt and feature audit. Table 56 reports the 720-case study; Figures 16–20 resolve input-to-output attribution. At the same measurement budget, valid outputs can therefore differ sharply in their physical relevance.

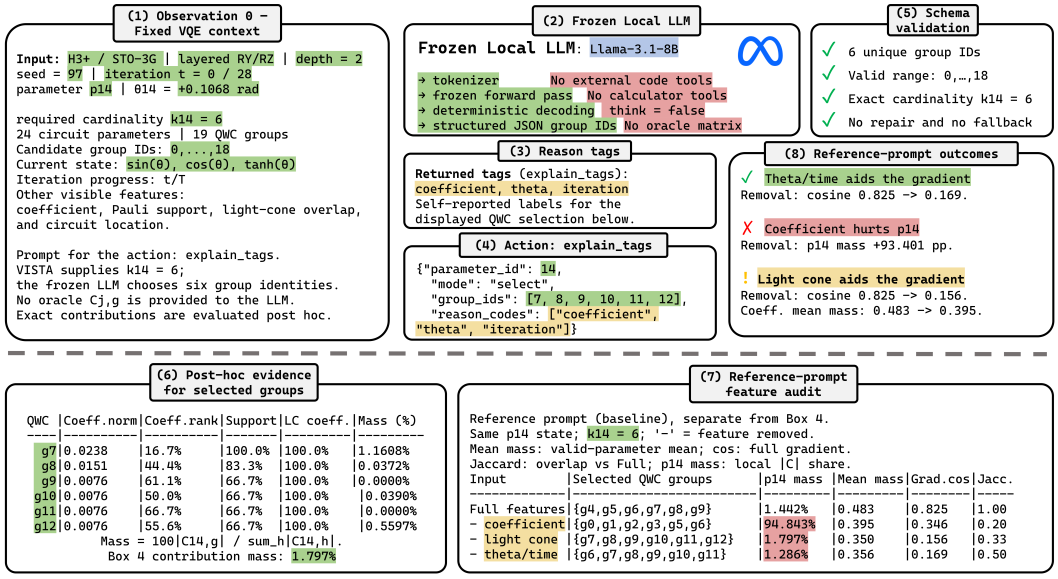

Figure 5: **Llama-3.1-8B observable decision trace.** On the shared H_3^+ snapshot, the explanation-prompt action captures 1.797% of the exact p_{14} contribution mass. Under the separate reference prompt, deleting coefficient fields raises local mass from 1.442% to 94.843%, identifying a harmful coefficient effect on that reference decision. Box (8) reports reference-prompt feature utility; prompt-specific actions and metrics are documented in Appendix N.4.

How to read the trace. The upper half records the model-visible input, frozen inference, returned IDs, and schema validation under the explanation prompt. The lower half attaches exact contributions and probes feature utility under the reference prompt. These views answer complementary questions: *what action the model returns, how much physical relevance that action retains, and how the reference selection responds to changes in its inputs.*

Observed decision. Llama returns six unique, in-range IDs, $\{g_7, \ldots, g_{12}\}$, with the requested cardinality and a valid schema. The action is directly executable, yet its captured mass is 1.797%. Because the measurement count is already fixed, the scientific distinction lies in the identities of the measured groups. The example makes the learned ranking task concrete: the controller must allocate the same number of measurements to the groups carrying the parameter's actual contribution.

Physical information does not guarantee effective selection

- Physical descriptors can help or mislead the same LLM: removing coefficients improves Llama's local selection but reduces its average contribution coverage.
- Removing light-cone or trajectory cues sharply degrades gradient alignment, showing that their importance extends beyond the contribution captured for a single parameter.

Task-specific relevance guides scientific decisions

- Our controller learns which measurement groups matter for each parameter. Matched-budget comparisons support the value of this learned ranking for energy–cost performance.
- These findings motivate expert-guided LLM workflows that prioritize task-relevant physical information and validate decisions through their scientific outcomes.

From one decision to closed-loop performance. Captured mass measures absolute contribution coverage for one parameter, while gradient cosine assesses alignment with the full gradient. Closed-loop evaluations assess the cumulative impact of repeated selections on final energy and measurement cost, extending the local trace analysis to complete optimization trajectories.

Across the feature audit. Over the paired audit, deleting coefficients or light-cone features reduces captured mass by 6.4 and 4.9 percentage points (Figure 22). Figure 21 relates these reference-prompt effects to separately reported tags. Positive and negative deletion effects distinguish beneficial and harmful feature utility, with model-specific responses to the same physical descriptors.

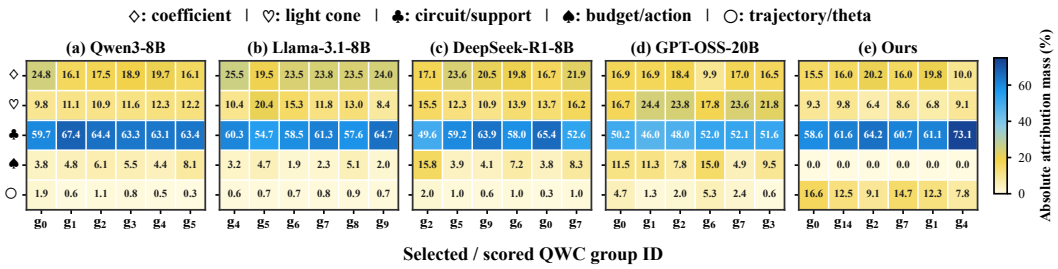

Figure 6: **Feature-family attribution on a shared $\mathbf{H}_3^+$ snapshot.** Panels (a)–(d): frozen LLMs; (e): ours. Cells show within-model absolute IG mass. Despite receiving the full input feature set, frozen LLMs do not automatically make substantial use of trajectory information without targeted intervention: it ranks last throughout the numerical audit (Appendix N.6). Our controller assigns 12.16% to current-state features. Scores and baselines: Appendix N.5; full maps: Appendix Figures 16–20.

The learned controller uses optimization state. Our controller combines circuit/support information (63.24%) with a clear current-state contribution (12.16%) in Figure 6. Trajectory denotes $[\sin\theta_j, \cos\theta_j, \tanh\theta_j, t/T]$: the angle and normalized optimization step. Native-checkpoint attribution identifies their explicit use in QWC ranking, combining observable structure with the circuit's current state. The fixed-snapshot audit in Appendix N.6 corroborates consistently low trajectory attribution in all four frozen LLMs across quadrature and precision settings.

Context capacity and ranking quality play different roles. In Table 54, increasing capacity from 8K to 16K and 32K raises protocol success from 60.4% to 81.2% and 97.9%; valid-case captured mass is 0.562, 0.564, and 0.554. A minimal prompt reaches 100% success at 140.6k prompt tokens per case, versus 160.2k for the domain reference. Direct-ID output improves paired mass by 3.5 percentage points at 95.8% success (Table 56). Thus context capacity improves executable coverage, while physical ranking determines selection quality. The best tested GPT-OSS prompt yields $\Delta E = +0.5056 \pm 0.6562$ Ha at 52.4% saving online (Appendix Tables 55 and 57).

Long-horizon validation. Appendix L, Table 62, extends quantum optimization to 400 Adam steps with simulated shot noise and frozen predictor weights. All 15 molecule–seed pairs reach within 1.6 mHa of their attainable exact-Adam reference at a checkpoint. Each molecule's mean final energy stays within 0.18 mHa of Full-QWC, with 24.7–92.1% shot savings. FCI anchors the absolute energy scale; the paired Full-QWC comparison isolates sparse measurement under the same ansatz. Savings persist across the long-horizon optimization.

Takeaway. Figure 6 suggests that the learned controller better exploits trajectory information, while the tested frozen LLMs struggle to prioritize these cues without explicit human guidance. This motivates expert-guided scientific workflows that help LLMs recognize and effectively use information critical to the task. This raises an interesting question: *What happens if we explicitly guide LLMs to use trajectory information?* We will explore this question across a broader range of quantum tasks in future work.

7 Conclusion

In this work, we uncover parameter-specific sparsity in QWC contributions to variational quantum gradients and develop a compact learned controller that exploits this structure for adjustable accuracy–cost trade-offs. Our evaluation of LLMs without fine-tuning reveals both capabilities and limitations: frontier models with retrieved training examples reach millihartree-scale gaps to Full-QWC, while local-model audits reveal uneven utility of the supplied physical information. These findings motivate AI-assisted scientific workflows in which human domain expertise guides problem representation, task objectives, and physical validation. By characterizing model capabilities and failure modes in a concrete scientific task, this work provides an important first step toward developing and evaluating such expert-guided workflows.

AI USE STATEMENT

In this work, we used generative AI tools for research ideation and feedback on methodology and experimental design, assistance with implementation, literature discovery, and editing the manuscript for clarity. The LLMs evaluated as measurement-selection baselines and their protocols are described separately in the paper. The authors reviewed AI-assisted material, checked reported numerical claims against experimental artifacts, and take full responsibility for the final text, claims, and research artifacts.

ETHICS STATEMENT

This study evaluates measurement selection in simulated quantum optimization using computationally generated Hamiltonians and structured physical features. It does not involve human participants or personal data.

REPRODUCIBILITY STATEMENT

Appendices M–N specify implementations, metrics, and protocol provenance; preceding appendices retain all experimental results.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Christopher J Anders, Kim A Nicoli, Bingting Wu, Naima Elosegui, Samuele Pedrielli, Lena Funcke, Karl Jansen, Stefan Kühn, and Shinichi Nakajima. Adaptive observation cost control for variational quantum eigensolvers. *arXiv preprint arXiv:2502.01704*, 2025.

Anthropic. Anthropic's transparency hub, 2026. URL https://www.anthropic.com/transparency.

Andrew Arrasmith, Lukasz Cincio, Rolando D Somma, and Patrick J Coles. Operator sampling for shot-frugal optimization in variational algorithms. *arXiv preprint arXiv:2004.06252*, 2020.

Alba Cervera-Lierta, Jakob S Kottmann, and Alán Aspuru-Guzik. Meta-variational quantum eigensolver: Learning energy profiles of parameterized hamiltonians for quantum simulation. *PRX Quantum*, 2(2):020329, 2021.

Ophelia Crawford, Barnaby Van Straaten, Daochen Wang, Thomas Parks, Earl Campbell, and Stephen Brierley. Efficient quantum measurement of pauli operators in the presence of finite sampling error. *Quantum*, 5:385, 2021.

Guillermo García-Pérez, Matteo AC Rossi, Boris Sokolov, Francesco Tacchino, Panagiotis Kl Barkoutsos, Guglielmo Mazzola, Ivano Tavernelli, and Sabrina Maniscalco. Learning to measure: Adaptive informationally complete generalized measurements for quantum algorithms. *Prx quantum*, 2(4):040342, 2021.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.

Pranav Gokhale, Olivia Angiuli, Yongshan Ding, Kaiwen Gui, Teague Tomesh, Martin Suchara, Margaret Martonosi, and Frederic T Chong. Minimizing state preparations in variational quantum eigensolver by partitioning into commuting families. *arXiv preprint arXiv:1907.13623*, 2019.

Jérôme F Gonthier, Maxwell D Radin, Corneliu Buda, Eric J Doskocil, Clena M Abuan, and Jhonathan Romero. Measurements as a roadblock to near-term practical quantum advantage in chemistry: Resource analysis. *Physical Review Research*, 4(3):033154, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Harper R Grimsley, Sophia E Economou, Edwin Barnes, and Nicholas J Mayhall. An adaptive variational algorithm for exact molecular simulations on a quantum computer. *Nature communications*, 10(1):3007, 2019.

Andi Gu, Angus Lowe, Pavel A Dub, Patrick J Coles, and Andrew Arrasmith. Adaptive shot allocation for fast convergence in variational quantum algorithms. *arXiv preprint arXiv:2108.10434*, 2021.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Charles Hadfield, Sergey Bravyi, Rudy Raymond, and Antonio Mezzacapo. Measurements of quantum hamiltonians with locally-biased classical shadows. *Communications in Mathematical Physics*, 391(3):951–967, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Warren J Hehre, Robert F Stewart, and John A Pople. Self-consistent molecular-orbital methods. i. use of gaussian expansions of slater-type atomic orbitals. *The Journal of Chemical Physics*, 51 (6):2657–2664, 1969.

Hsin-Yuan Huang, Richard Kueng, and John Preskill. Predicting many properties of a quantum system from very few measurements. *Nature Physics*, 16(10):1050–1057, 2020.

IBM Quantum. Fake provider: qiskit_ibm_runtime.fake_provider, 2026a. URL https://qiskit.github.io/qiskit-ibm-runtime/apidocs/fake_provider.html. Accessed: 2026-07-17.

IBM Quantum. Build noise models, 2026b. URL https://quantum.cloud.ibm.com/docs/en/guides/build-noise-models. Accessed: 2026-07-17.

Azhar Ikhtiarudin, Gagus Ketut Sunnardianto, Fadjar Fathurrahman, Mohammad Kemal Agusta, and Hermawan Kresno Dipojono. Shot-efficient adapt-vqe via reused pauli measurements and variance-based shot allocation. *arXiv preprint arXiv:2507.16879*, 2025.

Artur F Izmaylov, Tzu-Ching Yen, Robert A Lang, and Vladyslav Verteletskyi. Unitary partitioning approach to the measurement problem in the variational quantum eigensolver method. *Journal of chemical theory and computation*, 16(1):190–195, 2020.

Ali Javadi-Abhari, Matthew Treinish, Kevin Krsulich, Christopher J Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D Nation, Lev S Bishop, Andrew W Cross, et al. Quantum computing with qiskit. *arXiv preprint arXiv:2405.08810*, 2024.

Abhinav Kandala, Antonio Mezzacapo, Kristan Temme, Maika Takita, Markus Brink, Jerry M Chow, and Jay M Gambetta. Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *nature*, 549(7671):242–246, 2017.

Jonas M Kübler, Andrew Arrasmith, Lukasz Cincio, and Patrick J Coles. An adaptive optimizer for measurement-frugal variational algorithms. *Quantum*, 4:263, 2020.

Senwei Liang, Linghua Zhu, Xiaolin Liu, Chao Yang, and Xiaosong Li. Artificial-intelligence-driven shot reduction in quantum measurement. *Chemical Physics Reviews*, 5(4), 2024.

Sam McArdle, Suguru Endo, Alán Aspuru-Guzik, Simon C Benjamin, and Xiao Yuan. Quantum computational chemistry. *Reviews of Modern Physics*, 92(1):015003, 2020.

Jarrod R McClean, Jonathan Romero, Ryan Babbush, and Alán Aspuru-Guzik. The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2):023023, 2016.

Jarrod R McClean, Sergio Boixo, Vadim N Smelyanskiy, Ryan Babbush, and Hartmut Neven. Barren plateaus in quantum neural network training landscapes. *Nature communications*, 9(1):4812, 2018.

Jarrod R McClean, Nicholas C Rubin, Kevin J Sung, Ian D Kivlichan, Xavier Bonet-Monroig, Yudong Cao, Chengyu Dai, E Schuyler Fried, Craig Gidney, Brendan Gimby, et al. Openfermion: the electronic structure package for quantum computers. *Quantum Science & Technology*, 5(3): 034014, 2020.

OpenAI. All models, 2026. URL https://developers.openai.com/api/docs/models/all.

Yifeng Peng, Xinyi Li, Samuel Yen-Chi Chen, Kaining Zhang, Zhiding Liang, Ying Wang, and Yuxuan Du. Titan: A trajectory-informed technique for adaptive parameter freezing in large-scale vqe. *Advances in Neural Information Processing Systems*, 38:47294–47315, 2026.

Alberto Peruzzo, Jarrod McClean, Peter Shadbolt, Man-Hong Yung, Xiao-Qi Zhou, Peter J Love, Alán Aspuru-Guzik, and Jeremy L O'brien. A variational eigenvalue solver on a photonic quantum processor. *Nature communications*, 5(1):4213, 2014.

Maria Schuld, Ville Bergholm, Christian Gogolin, Josh Izaac, and Nathan Killoran. Evaluating analytic gradients on quantum hardware. *Physical Review A*, 99(3):032331, 2019.

Jacob T Seeley, Martin J Richard, and Peter J Love. The bravyi-kitaev transformation for quantum computation of electronic structure. *The Journal of chemical physics*, 137(22), 2012.

Qiming Sun, Timothy C Berkelbach, Nick S Blunt, George H Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D McClain, Elvira R Sayfutyarova, Sandeep Sharma, et al. Pyscf: the python-based simulations of chemistry framework. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1):e1340, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vladyslav Verteletskyi, Tzu-Ching Yen, and Artur F Izmaylov. Measurement optimization in the variational quantum eigensolver using a minimum clique cover. *The Journal of chemical physics*, 152(12), 2020.

David Wierichs, Josh Izaac, Cody Wang, and Cedric Yen-Yu Lin. General parameter-shift rules for quantum gradients. *Quantum*, 6:677, 2022.

William K Wootters and Wojciech H Zurek. A single quantum cannot be cloned. *Nature*, 299(5886): 802–803, 1982.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Tzu-Ching Yen, Vladyslav Verteletskyi, and Artur F Izmaylov. Measuring all compatible operators in one series of single-qubit measurements using unitary transformations. *Journal of chemical theory and computation*, 16(4):2400–2409, 2020.

Andrew Zhao, Andrew Tranter, William M Kirby, Shu Fay Ung, Akimasa Miyake, and Peter J Love. Measurement reduction in variational quantum algorithms. *Physical Review A*, 101(6):062322, 2020.

Linghua Zhu, Senwei Liang, Chao Yang, and Xiaosong Li. Optimizing shot assignment in variational quantum eigensolver measurement. *Journal of Chemical Theory and Computation*, 20(6): 2390–2403, 2024.

A Extended Related Work

Measurement-frugal VQE. Commuting partitioning, unitary partitioning, and compatible-observable measurement reduce the number of bases used to estimate the Hamiltonian (Zhao et al., 2020; Izmaylov et al., 2020; Yen et al., 2020); classical-shadow variants instead reuse randomized measurements across many observables (Huang et al., 2020; Hadfield et al., 2022). A complementary line allocates a finite shot budget according to gradient or observable variance (Kübler et al., 2020; Gu et al., 2021; Zhu et al., 2024), while operator sampling constructs stochastic gradient estimates from selected Hamiltonian components (Arrasmith et al., 2020). These approaches optimize the global grouping, the number of shots, or a sampled operator distribution. Our method operates after a QWC partition is fixed and asks which groups should be retained for each individual parameter.

Learned measurement adaptation. Learning has also been used to adapt generalized measurements and reduce observation cost (García-Pérez et al., 2021; Liang et al., 2024; Anders et al., 2025). Our output space is different: it is a variable-size $P \times G$ relevance matrix tied directly to parameter-shift gradient decomposition. To our knowledge, prior work does not jointly learn this parameter-by-QWC grid and charge cost at the physical measurement-setting level.

Ansatz adaptation and parameter freezing. Ansatz adaptation and parameter freezing reduce optimization cost along an orthogonal axis. ADAPT-VQE grows an operator pool selectively (Grimsley et al., 2019; Ikhtiarudin et al., 2025), whereas TITAN predicts parameters that can be frozen along a trajectory (Peng et al., 2026). Meta-VQE learns across related Hamiltonians to accelerate energy-profile estimation (Cervera-Lierta et al., 2021). Our method neither changes the ansatz nor requires a parameter to be frozen: for every parameter that is evaluated, it sparsifies the QWC groups entering that gradient. We therefore include both TITAN and TITAN+**Ours** to separate and compose the two cost axes.

B Additional In-Distribution Results

B.1 Complete Main-Panel Baselines

Table 5: **Complete version of Table 1.** This table preserves the default-threshold TITAN (Peng et al., 2026) and TITAN+**Ours** rows, the cost-matched † rows, and all one-sided no-worse indicators omitted from the main table for readability. All molecules, seeds, online-VQE settings, and highlighting rules are unchanged.

Setting	Metric	Method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 1	ΔE vs Full ↓	**Ours**	+0.0011	+0.0015	+0.0005	−0.0006	+0.0023
Depth 1	ΔE vs Full ↓	Random	+0.2359	+1.2220	+0.8916	+1.4117	+3.4982
Depth 1	ΔE vs Full ↓	Global coeff.	+0.0038	+0.0019	+0.0122	+0.0089	+0.0016
Depth 1	ΔE vs Full ↓	TITAN (Peng et al.)	+0.0228	+0.1040	+0.0871	+0.0220	−0.0881
Depth 1	ΔE vs Full ↓	TITAN+**Ours**	+0.0227	+0.1040	+0.0869	+0.0218	−0.0875
Depth 1	ΔE vs Full ↓	TITAN†	+0.0128	+0.1299	+0.0772	−0.0642	−0.0648
Depth 1	ΔE vs Full ↓	TITAN†+**Ours**	+0.0127	+0.1297	+0.0776	−0.0649	−0.0646
Depth 1	Shot gain ↑	**Ours**	40.4%	58.9%	63.0%	74.7%	86.9%
Depth 1	Shot gain ↑	Random	42.0%	60.7%	66.5%	79.2%	90.2%
Depth 1	Shot gain ↑	Global coeff.	41.7%	60.0%	62.9%	75.0%	87.0%
Depth 1	Shot gain ↑	TITAN (Peng et al.)	61.8%	56.2%	58.2%	62.7%	71.9%
Depth 1	Shot gain ↑	TITAN+**Ours**	**80.0%**	90.5%	93.5%	96.1%	99.7%
Depth 1	Shot gain ↑	TITAN†	43.8%	59.4%	62.7%	70.3%	81.8%
Depth 1	Shot gain ↑	TITAN†+**Ours**	66.8%	**91.7%**	**95.0%**	**97.9%**	**99.9%**
Depth 1	No-worse vs Full	**Ours**	✓	✓	✓	✓	✗
Depth 1	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 1	No-worse vs Full	Global coeff.	✗	✗	✗	✗	✗
Depth 1	No-worse vs Full	TITAN (Peng et al.)	✗	✗	✗	✗	✓
Depth 1	No-worse vs Full	TITAN+**Ours**	✗	✗	✗	✗	✓
Depth 1	No-worse vs Full	TITAN†	✗	✗	✗	✓	✓
Depth 1	No-worse vs Full	TITAN†+**Ours**	✗	✗	✗	✓	✓
Depth 2	ΔE vs Full ↓	**Ours**	+0.0003	−0.0011	−0.0015	+0.0158	−0.0049
Depth 2	ΔE vs Full ↓	Random	+0.1656	+0.4704	+0.6527	+0.9343	+4.0043
Depth 2	ΔE vs Full ↓	Global coeff.	+0.0015	−0.0004	+0.0233	+0.0379	−0.0049
Depth 2	ΔE vs Full ↓	TITAN (Peng et al.)	−0.0200	−0.1194	−0.0031	+0.1521	+0.0806
Depth 2	ΔE vs Full ↓	TITAN+**Ours**	−0.0200	−0.1200	−0.0032	+0.1538	+0.0784
Depth 2	ΔE vs Full ↓	TITAN†	+0.0075	+0.0116	+0.0153	+0.0472	+0.2903
Depth 2	ΔE vs Full ↓	TITAN†+**Ours**	+0.0072	+0.0129	+0.0140	+0.0493	+0.2920
Depth 2	Shot gain ↑	**Ours**	25.6%	31.6%	44.9%	51.8%	82.6%
Depth 2	Shot gain ↑	Random	26.3%	33.2%	47.7%	53.7%	83.4%
Depth 2	Shot gain ↑	Global coeff.	25.7%	31.9%	44.7%	51.7%	82.5%
Depth 2	Shot gain ↑	TITAN (Peng et al.)	54.0%	54.4%	50.0%	59.1%	77.1%
Depth 2	Shot gain ↑	TITAN+**Ours**	**70.9%**	**75.7%**	**79.9%**	**86.1%**	98.7%
Depth 2	Shot gain ↑	TITAN†	23.8%	39.0%	41.9%	46.8%	82.7%
Depth 2	Shot gain ↑	TITAN†+**Ours**	40.2%	64.0%	74.5%	79.0%	**99.3%**
Depth 2	No-worse vs Full	**Ours**	✓	✓	✓	✗	✓
Depth 2	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 2	No-worse vs Full	Global coeff.	✓	✓	✗	✗	✓
Depth 2	No-worse vs Full	TITAN (Peng et al.)	✓	✓	✓	✗	✗
Depth 2	No-worse vs Full	TITAN+**Ours**	✓	✓	✓	✗	✗
Depth 2	No-worse vs Full	TITAN†	✗	✗	✗	✗	✗
Depth 2	No-worse vs Full	TITAN†+**Ours**	✗	✗	✗	✗	✗

B.2 In-Distribution Aggregate and Budget Details

Table 6: **Depth-wise aggregate at $\tau = 0.995$.** Mean $|\Delta E|$ averages seed-level absolute gaps over four seeds and five molecules at each depth. No-worse marks whether the aggregate mean signed ΔE is at most 1.6 mHa above Full.

Setting	Mean $\lvert\Delta E\rvert \downarrow$	No-worse	Mean shot gain ↑
Full-QWC reference	0.00000	✓	0.0%
Ours depth 1	0.00150	✓	**64.8%**
Ours depth 2	0.00978	✗	**47.3%**
Ours depth 3	0.00571	✓	**36.7%**
Random depth 1	1.45188	✗	**67.7%**
Random depth 2	1.24548	✗	**48.9%**
Random depth 3	0.82189	✗	**37.3%**
Ours all	0.00567	✓	**49.6%**
Random all	1.17308	✗	**51.3%**

Table 7: **Detailed budget usage at $\tau = 0.995$.** All molecular Hamiltonians use the STO-3G minimal basis; molecule labels report the resulting qubit counts. Groups/param. reports the mean selected QWC groups per parameter, shot gain is relative to full $2PGS$, and No-worse vs Full uses the one-sided 1.6 mHa criterion ($\Delta E \leq 1.6$ mHa; lower energy is accepted). Boldface marks the better **Ours**-vs-Random entry for each molecule-depth comparison.

Molecule	Method	Full QWC	Groups / param. ↓			Shot gain ↑			No-worse vs Full			Best $\Delta E \downarrow$
		groups/param.	D1	D2	D3	D1	D2	D3	D1	D2	D3	
H2 (4q)	**Ours**	5	2.98	3.72	**4.09**	**40.4%**	**25.6%**	**18.1%**	✓	✓	✓	+0.0003
H2 (4q)	Random	5	**2.90**	**3.68**	4.11	**42.0%**	**26.3%**	**17.9%**	✗	✗	✓	−0.0145
HeH+ (4q)	**Ours**	11	4.52	7.53	**8.26**	**58.9%**	**31.6%**	**24.9%**	✓	✓	✗	−0.0011
HeH+ (4q)	Random	11	**4.32**	**7.35**	8.30	**60.7%**	**33.2%**	**24.6%**	✗	✗	✗	+0.0513
H3+ (6q)	**Ours**	19	7.03	10.47	13.33	**63.0%**	**44.9%**	**29.9%**	✓	✓	✓	+0.0005
H3+ (6q)	Random	19	**6.37**	**9.94**	**13.31**	**66.5%**	**47.7%**	**29.9%**	✗	✗	✗	+0.2864
H4 chain (8q)	**Ours**	75	18.96	36.15	48.60	**74.7%**	**51.8%**	**35.2%**	✓	✗	✓	−0.0006
H4 chain (8q)	Random	75	**15.61**	**34.71**	**47.43**	**79.2%**	**53.7%**	**36.8%**	✗	✗	✗	+0.2903
LiH (12q)	**Ours**	179	23.40	31.16	44.09	**86.9%**	**82.6%**	**75.4%**	✗	✓	✓	−0.0017
LiH (12q)	Random	179	**17.57**	**29.78**	**40.82**	**90.2%**	**83.4%**	**77.2%**	✗	✗	✗	+3.3113

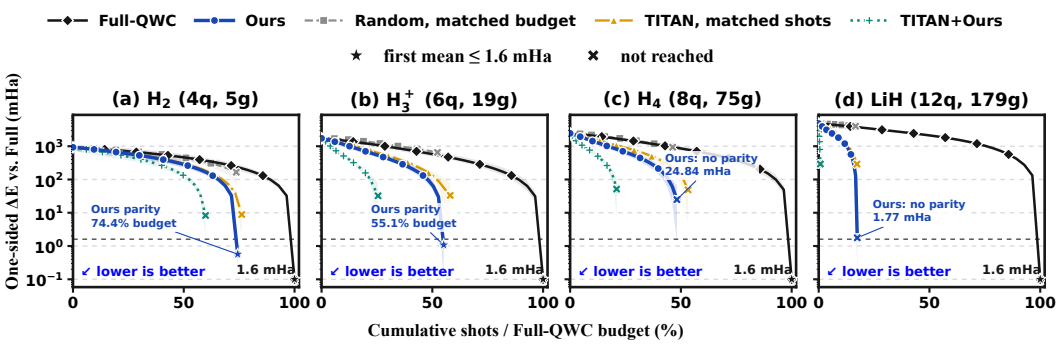

Figure 7: **Shots-to-chemical-parity dynamics at depth 2.** Curves report the mean across four paired seeds for H2, H3+, an H4 chain, and LiH; shaded regions are 95% bootstrap confidence intervals. The horizontal axis is cumulative parameter-shift shots normalized by each seed's Full-QWC budget. The vertical axis is the one-sided energy excess above the paired Full-QWC final energy, so lower energies incur no penalty; the dashed line marks 1.6 mHa chemical parity. Stars mark the first optimizer checkpoint where the four-seed mean reaches parity, while crosses mark methods that never reach it. Our method first reaches parity at 74.4% of the Full-QWC budget on H2 and 55.1% on H3+; it does not reach parity on H4 and finishes narrowly above it on LiH (1.77 mHa at 17.4% budget). Unlike Table 1, which compares final endpoints, this figure isolates convergence efficiency and shots-to-target.

B.3 Shots-to-Chemical-Parity Dynamics

Figure 7 complements the endpoint comparison with cumulative-shot convergence trajectories.

Post-hoc budget control. We vary only the cumulative top-mass threshold τ after training; model weights, labels, and online oracle access remain unchanged. One selector trained on depths 1–4 is evaluated on unseen depths 5–7. Increasing τ from 0.98 to 0.997 reduces the proposed method's mean absolute energy error from 22.5 to 7.2 mHa while retaining 16.2% shot saving. At closely matched budgets, Random remains at 264–606 mHa. The result exposes a controllable accuracy–measurement frontier without retraining.

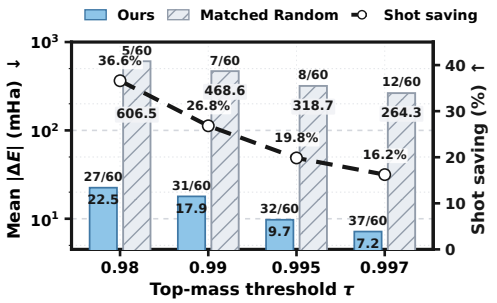

Figure 8: **Depth-OOD threshold Pareto frontier.** Bars show mean $|\Delta E|$ to Full QWC; labels above them count no-worse trajectories out of 60 under the one-sided 1.6 mHa criterion. The line shows the proposed method's shot saving; Random uses an almost identical budget. Full values are in Appendix C, Table 8.

C Threshold Control and Depth-OOD Pareto Details

C.1 Depth-OOD Pareto Visualization

C.2 Complete Numerical Threshold Frontier

Table 8: **Complete numerical counterpart of Figure 8.** A single model trained on depths 1–4 is reused across thresholds and evaluated on unseen depths 5–7. Only the inference top-mass threshold changes; energy gaps are in Ha and costs are nominal. This duplicates the depth-OOD aggregate in Table 10, not an independent replication. No-worse seeds have final energy at most 1.6 mHa above Full QWC; lower final energy is accepted. Further threshold, depth-wise, and molecule-wise results appear in Tables 10, 14, and 15.

Threshold	**Ours**					Random same budget		
τ	Mean $\lvert\Delta E\rvert \downarrow$	Mean ΔE	No-worse seeds	Shot gain ↑	Groups/param.	Mean $\lvert\Delta E\rvert \downarrow$	No-worse seeds	Shot gain ↑
0.98	0.02246	+0.0066	27/60	**36.6%**	24.43	0.60645	5/60	**37.0%**
0.99	0.01791	+0.0025	31/60	**26.8%**	30.47	0.46857	7/60	**27.4%**
0.995	0.00974	−0.0010	32/60	**19.8%**	35.67	0.31867	8/60	**20.4%**
0.997	0.00716	−0.0012	37/60	**16.2%**	38.90	0.26431	12/60	**16.7%**

C.3 Top-Mass Threshold Sweep Summary

The current same-family ansatz-OOD table uses $\tau = 0.997$. To make the budget dial explicit, Tables 9 and 10 summarize the existing threshold sweeps. In all cases, τ is the cumulative top-mass threshold used to convert the predicted parameter-by-QWC relevance matrix into selected QWC groups:

$$\sum_{k \in S_j(\tau)} \widehat{w}_{j,k} \geq \tau. \tag{8}$$

For a fixed predicted row, larger τ retains a nested set of groups. This does not guarantee monotone final energy or closed-loop average cost, and the historical points below are not all evaluations of one checkpoint.

Table 9: **Historical held-out-seed operating points, not one checkpoint's sweep.** Depths 1–3 use exact-expectation VQE and nominal shot accounting. Earlier 0.95/0.98/0.99 points belong to separately trained models; those models used the auxiliary count head. Only 0.995/0.997 reuse the $\tau_{\text{train}} = 0.99$ checkpoint with top-mass post-processing. Energy gaps are in Ha. Checkpoint and selection-mode provenance are in Appendix N.2; detailed results follow.

Threshold	**Ours**				Random same budget		
τ	Mean $\lvert\Delta E\rvert \downarrow$	Mean ΔE	Shot gain ↑	Groups/param.	Mean $\lvert\Delta E\rvert \downarrow$	Shot gain ↑	Groups/param.
0.95	0.00917	+0.0073	**62.3%**	12.19	1.33512	**63.4%**	11.24
0.98	0.00984	+0.0040	**52.8%**	17.59	1.07202	**53.8%**	16.63
0.99	0.00439	**+0.0011**	**45.5%**	21.54	0.93589	**46.5%**	20.32
0.995	0.00567	**+0.0014**	**49.6%**	17.62	1.17308	**51.3%**	16.41
0.997	0.00352	**+0.0000**	**45.5%**	19.84	1.10026	**47.4%**	18.59

Table 10: **Depth-OOD threshold sweep.** A single depth-OOD model is trained on depths 1–4 and evaluated on unseen depths 5–7. Only the inference top-mass threshold changes; energy gaps are in Ha and shot costs are nominal. This differs from the historical table above (Appendix N.2). No-worse seeds count trajectories whose final energy is not more than 1.6 mHa above Full-QWC; lower energy is accepted.

Threshold	**Ours**					Random same budget		
τ	Mean $\lvert\Delta E\rvert \downarrow$	Mean ΔE	No-worse seeds	Shot gain ↑	Groups/param.	Mean $\lvert\Delta E\rvert \downarrow$	No-worse seeds	Shot gain ↑
0.98	0.02246	+0.0066	27/60	**36.6%**	24.43	0.60645	5/60	**37.0%**
0.99	0.01791	+0.0025	31/60	**26.8%**	30.47	0.46857	7/60	**27.4%**
0.995	0.00974	**-0.0010**	32/60	**19.8%**	35.67	0.31867	8/60	**20.4%**
0.997	0.00716	**-0.0012**	37/60	**16.2%**	38.90	0.26431	12/60	**16.7%**

C.4 STRICTER POINT: $\tau = 0.997$

The $\tau = 0.997$ point uses the same saved relevance model and increases only the top-mass post-processing threshold. It is intended as a conservative operating point for cases where $\tau = 0.995$ is close but not always within the no-worse criterion.

Table 11: **Matched-budget online VQE baseline comparison at $\tau = 0.997$.** The same saved relevance model is reused; only the top-mass post-processing threshold is increased to $\tau = 0.997$. For each depth and metric, our method is paired with a Random same-budget baseline that keeps the same number of selected QWC groups per parameter but chooses groups uniformly at random. Boldface marks the better method for each molecule-specific comparison.

Setting	Metric	Method	H2/STO-3G	HeH+/STO-3G	H3+/STO-3G	H4 chain/STO-3G	LiH/STO-3G
Depth 1	ΔE vs Full ↓	**Ours**	+0.0005	+0.0001	−0.0002	−0.0013	+0.0020
Depth 1	ΔE vs Full ↓	Random	+0.2192	+1.3164	+0.7630	+1.2211	+3.5105
Depth 1	Shot gain ↑	**Ours**	**37.2%**	**55.1%**	**59.5%**	**71.7%**	**84.9%**
Depth 1	Shot gain ↑	Random	**41.3%**	**56.6%**	**62.7%**	**75.5%**	**88.9%**
Depth 1	No-worse vs Full	**Ours**	✓	✓	✓	✓	✗
Depth 1	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 2	ΔE vs Full ↓	**Ours**	−0.0001	−0.0024	−0.0030	+0.0017	−0.0021
Depth 2	ΔE vs Full ↓	Random	+0.1092	+0.3799	+0.5018	+0.7784	+3.7420
Depth 2	Shot gain ↑	**Ours**	**21.8%**	**27.3%**	**39.8%**	**46.6%**	**79.1%**
Depth 2	Shot gain ↑	Random	**23.0%**	**29.3%**	**42.3%**	**47.8%**	**80.0%**
Depth 2	No-worse vs Full	**Ours**	✓	✓	✓	✗	✓
Depth 2	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 3	ΔE vs Full ↓	**Ours**	+0.0007	+0.0074	−0.0023	+0.0004	−0.0010
Depth 3	ΔE vs Full ↓	Random	+0.0092	+0.1772	+0.1448	+0.1285	+3.3875
Depth 3	Shot gain ↑	**Ours**	**14.8%**	**20.3%**	**24.7%**	**30.0%**	**70.4%**
Depth 3	Shot gain ↑	Random	**14.9%**	**20.6%**	**25.2%**	**30.7%**	**72.6%**
Depth 3	No-worse vs Full	**Ours**	✓	✗	✓	✓	✓
Depth 3	No-worse vs Full	Random	✗	✗	✗	✗	✗

Table 12: **Depth-wise aggregate at $\tau = 0.997$.** Mean $|\Delta E|$ averages the five molecule columns at each ansatz depth. No-worse marks whether the aggregate mean signed ΔE is at most 1.6 mHa above Full.

Setting	Mean $\lvert\Delta E\rvert$ ↓	No-worse	Mean shot gain ↑
Full-QWC reference	0.00000	✓	0.0%
Ours depth 1	0.00112	✓	**61.7%**
Ours depth 2	0.00448	✓	**42.9%**
Ours depth 3	0.00497	✓	**32.0%**
Random depth 1	1.40602	✗	**65.0%**
Random depth 2	1.10226	✗	**44.5%**
Random depth 3	0.79248	✗	**32.8%**
Ours all	0.00352	✓	**45.5%**
Random all	1.10026	✗	**47.4%**

Table 13: **Detailed budget usage at $\tau = 0.997$.** Groups/param. reports the mean selected QWC groups per parameter, shot gain is relative to full $2PGS$, and No-worse vs Full uses the one-sided 1.6 mHa criterion ($\Delta E \leq 1.6$ mHa; lower energy is accepted). Boldface marks the better our method-vs-Random entry for each molecule-depth comparison.

Molecule	Method	Full QWC	Groups / param. ↓			Shot gain ↑			No-worse vs Full			Best ΔE ↓
		groups/param.	D1	D2	D3	D1	D2	D3	D1	D2	D3	
H2/STO-3G	**Ours**	5	3.14	3.91	4.26	**37.2%**	**21.8%**	**14.8%**	✓	✓	✓	−0.0001
H2/STO-3G	Random	5	**2.94**	**3.85**	**4.26**	**41.3%**	**23.0%**	**14.9%**	✗	✗	✗	+0.0092
HeH+/STO-3G	**Ours**	11	4.94	8.00	8.77	**55.1%**	**27.3%**	**20.3%**	✓	✓	✗	+0.0001
HeH+/STO-3G	Random	11	**4.77**	**7.78**	**8.74**	**56.6%**	**29.3%**	**20.6%**	✗	✗	✗	+0.1772
H3+/STO-3G	**Ours**	19	7.69	11.44	14.31	**59.5%**	**39.8%**	**24.7%**	✓	✓	✓	−0.0002
H3+/STO-3G	Random	19	**7.08**	**10.97**	**14.20**	**62.7%**	**42.3%**	**25.2%**	✗	✗	✗	+0.1448
H4 chain/STO-3G	**Ours**	75	21.26	40.03	52.52	**71.7%**	**46.6%**	**30.0%**	✓	✗	✓	+0.0004
H4 chain/STO-3G	Random	75	**18.40**	**39.14**	**51.98**	**75.5%**	**47.8%**	**30.7%**	✗	✗	✗	+0.1285
LiH/STO-3G	**Ours**	179	26.99	37.46	52.92	**84.9%**	**79.1%**	**70.4%**	✗	✓	✓	−0.0010
LiH/STO-3G	Random	179	**19.89**	**35.75**	**49.06**	**88.9%**	**80.0%**	**72.6%**	✗	✗	✗	+3.3875

C.5 Depth-OOD Pareto Sweep

This sweep reuses the depth-OOD relevance model and changes only the top-mass threshold. This isolates the post-hoc budget dial: lower thresholds save more QWC measurements, while higher thresholds recover more of the Full-QWC gradient signal.

Table 14: **Depth-wise view of the OOD Pareto sweep.** A single Transformer relevance model is trained on ansatz depths 1–4 and evaluated without retraining on unseen depths 5–7. Each row pools the five PySCF/STO-3G molecular Hamiltonians (H2, HeH+, H3+, H4 chain, and LiH) and four held-out seeds at one unseen evaluation depth. Mean ΔE is reported as mean $\pm$ standard deviation over the 20 paired trajectories. CA marks whether the mean signed ΔE is no more than 1.6 mHa above the Full-QWC reference; lower final energy is accepted. Group gain is the relative reduction in selected QWC groups per parameter versus Full-QWC.

Threshold	Depth	**Ours**				Random same budget			
τ	d	Mean $\Delta E \pm$ s.d.	CA	Shot gain ↑	Group gain ↑	Mean $\Delta E \pm$ s.d.	CA	Shot gain ↑	Group gain ↑
0.98	5	$+0.0219 \pm 0.0618$	✗	**39.4%**	**62.3%**	$+0.7969 \pm 1.2005$	✗	**39.9%**	**64.4%**
0.98	6	-0.0054 ± 0.0243	✓	**36.5%**	**57.4%**	$+0.5774 \pm 0.9583$	✗	**36.9%**	**59.3%**
0.98	7	$+0.0035 \pm 0.0124$	✗	**33.9%**	**53.5%**	$+0.4024 \pm 0.7036$	✗	**34.1%**	**55.0%**
0.99	5	$+0.0048 \pm 0.0555$	✗	**29.7%**	**52.4%**	$+0.7167 \pm 1.1610$	✗	**30.4%**	**54.8%**
0.99	6	$+0.0026 \pm 0.0256$	✗	**26.7%**	**46.9%**	$+0.4179 \pm 0.6588$	✗	**27.2%**	**49.0%**
0.99	7	$+0.0002 \pm 0.0130$	✓	**24.1%**	**42.5%**	$+0.2481 \pm 0.4441$	✗	**24.6%**	**44.1%**
0.995	5	-0.0016 ± 0.0276	✓	**22.6%**	**43.8%**	$+0.5389 \pm 0.9643$	✗	**23.5%**	**46.2%**
0.995	6	-0.0014 ± 0.0090	✓	**19.6%**	**37.8%**	$+0.2580 \pm 0.4218$	✗	**20.1%**	**39.2%**
0.995	7	-0.0001 ± 0.0107	✓	**17.3%**	**33.2%**	$+0.1470 \pm 0.2903$	✗	**17.7%**	**34.5%**
0.997	5	-0.0025 ± 0.0155	✓	**18.9%**	**38.3%**	$+0.4968 \pm 0.8809$	✗	**19.5%**	**39.9%**
0.997	6	-0.0006 ± 0.0105	✓	**16.0%**	**32.1%**	$+0.1665 \pm 0.3072$	✗	**16.4%**	**33.4%**
0.997	7	-0.0005 ± 0.0098	✓	**13.8%**	**27.6%**	$+0.0916 \pm 0.2070$	✗	**14.1%**	**28.7%**

Table 15: **Molecule-wise OOD summary at $\tau = 0.997$.** This conservative point gives the best global mean $|\Delta E|$ in the sweep while still retaining nonzero shot savings, especially for richer Hamiltonians with many QWC groups.

Molecule	**Ours**			Random same budget		
	Mean $\lvert\Delta E\rvert$ ↓	Mean ΔE	Shot gain ↑	Mean $\lvert\Delta E\rvert$ ↓	Mean ΔE	Shot gain ↑
H2/STO-3G	0.00140	$+0.0003$	**4.8%**	0.01690	$+0.0066$	**4.7%**
HeH+/STO-3G	0.00735	-0.0024	**7.6%**	0.06000	$+0.0424$	**7.5%**
H3+/STO-3G	0.00668	-0.0036	**8.9%**	0.03097	$+0.0160$	**8.9%**
H4 chain/STO-3G	0.01298	-0.0011	**14.9%**	0.09763	$+0.0771$	**15.0%**
LiH/STO-3G	0.00742	$+0.0009$	**45.0%**	1.11605	$+1.1160$	**47.1%**

C.6 Previous Threshold Sweep Tables

The main paper reports the stricter post-processing point $\tau = 0.995$. For completeness, we keep the earlier threshold-sweep tables below. These tables use the same held-out-seed online VQE protocol and matched-budget random baseline, but reflect earlier trained operating points $\tau \in \{0.95, 0.98, 0.99\}$, including count-head selection rather than a single-checkpoint inference-threshold sweep. Appendix N.2 gives provenance. All energy gaps are in Ha and shot gains use nominal accounting.

C.7 Previous Threshold: $\tau = 0.98$

Table 16: **Matched-budget online VQE baseline comparison (Main Results: threshold $\tau = 0.98$).** The learned selector is trained on RY/RZ VQE trajectories from depths 1–3 using row-wise top-mass target threshold $\tau = 0.98$ and evaluated in closed-loop online VQE on held-out random seeds for the same five PySCF molecules and depths. For each depth and metric, our method is paired with a Random same-budget baseline that keeps the same predicted number of QWC groups per parameter but chooses groups uniformly at random. Boldface marks the better method for each molecule-specific comparison.

Setting	Metric	Method	H2/STO-3G	HeH+/STO-3G	H3+/STO-3G	H4 chain/STO-3G	LiH/STO-3G
Depth 1	ΔE vs Full ↓	**Ours**	−0.0000	+0.0028	+0.0040	−0.0021	+0.0012
Depth 1	ΔE vs Full ↓	Random	+0.2500	+1.3729	+0.7713	+1.1147	+3.2882
Depth 1	Shot gain ↑	**Ours**	**41.2%**	**59.2%**	**66.2%**	**73.7%**	**86.4%**
Depth 1	Shot gain ↑	Random	**43.7%**	**62.9%**	**69.0%**	**77.0%**	**89.8%**
Depth 1	No-worse vs Full	**Ours**	✓	✗	✗	✓	✓
Depth 1	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 2	ΔE vs Full ↓	**Ours**	+0.0244	+0.0103	−0.0021	+0.0214	−0.0032
Depth 2	ΔE vs Full ↓	Random	+0.1560	+0.4481	+0.4674	+0.8634	+3.4465
Depth 2	Shot gain ↑	**Ours**	**31.9%**	**40.5%**	**49.6%**	**56.3%**	**79.9%**
Depth 2	Shot gain ↑	Random	**30.8%**	**39.3%**	**49.7%**	**56.7%**	**81.2%**
Depth 2	No-worse vs Full	**Ours**	✗	✗	✓	✗	✓
Depth 2	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 3	ΔE vs Full ↓	**Ours**	−0.0022	+0.0121	−0.0026	−0.0020	−0.0014
Depth 3	ΔE vs Full ↓	Random	+0.0586	+0.1960	+0.2986	+0.2677	+3.0320
Depth 3	Shot gain ↑	**Ours**	**24.9%**	**31.3%**	**37.4%**	**41.8%**	**71.6%**
Depth 3	Shot gain ↑	Random	**24.0%**	**30.6%**	**36.7%**	**42.1%**	**73.0%**
Depth 3	No-worse vs Full	**Ours**	✓	✗	✓	✓	✓
Depth 3	No-worse vs Full	Random	✗	✗	✗	✗	✗

Table 16 shows that at threshold $\tau = 0.98$, our method reaches mean absolute final-energy gap 0.00984 while saving 52.8% of full-gradient measurement shots. The matched-budget random baseline has mean absolute gap 1.07202 with 53.8% shot saving, so the selected QWC groups matter beyond the budget alone.

Table 17: **Depth-wise aggregate (Main Results: threshold $\tau = 0.98$).** Same protocol as Table 16 with top-mass target threshold $\tau = 0.98$. Mean $|\Delta E|$ aggregates the five molecule columns at each ansatz depth; No-worse marks whether the aggregate mean signed ΔE is at most 1.6 mHa above Full.

Setting	Mean $\lvert\Delta E\rvert$ ↓	No-worse	Mean shot gain ↑
Full-QWC reference	0.00000	✓	0.0%
Ours depth 1	0.00217	✓	**65.3%**
Ours depth 2	0.01616	✗	**51.6%**
Ours depth 3	0.01119	✓	**41.4%**
Random depth 1	1.35940	✗	**68.5%**
Random depth 2	1.07628	✗	**51.5%**
Random depth 3	0.78037	✗	**41.3%**
Ours all	0.00984	✗	**52.8%**
Random all	1.07202	✗	**53.8%**

Table 18: **Detailed budget usage (Main Results: threshold $\tau = 0.98$).** Same table format with top-mass target threshold $\tau = 0.98$. Groups/param. reports the mean selected QWC groups per parameter, shot gain is relative to full $2PGS$, and No-worse vs Full uses the one-sided 1.6 mHa criterion ($\Delta E \leq 1.6$ mHa; lower energy is accepted). Boldface marks the better Ours-vs-Random entry for each molecule-depth comparison.

Molecule	Method	Full QWC	Groups / param. ↓			Shot gain ↑			No-worse vs Full			Best ΔE ↓
		groups/param.	D1	D2	D3	D1	D2	D3	D1	D2	D3	
H2/STO-3G	**Ours**	5	2.94	**3.41**	**3.75**	**41.2%**	**31.9%**	**24.9%**	✓	✗	✓	−0.0000
H2/STO-3G	Random	5	**2.82**	3.46	3.80	**43.7%**	**30.8%**	**24.0%**	✗	✗	✗	+0.0586
HeH+/STO-3G	**Ours**	11	4.49	**6.55**	**7.56**	**59.2%**	**40.5%**	**31.3%**	✗	✗	✗	+0.0028
HeH+/STO-3G	Random	11	**4.08**	6.68	7.63	**62.9%**	**39.3%**	**30.6%**	✗	✗	✗	+0.1960
H3+/STO-3G	**Ours**	19	6.42	9.58	**11.89**	**66.2%**	**49.6%**	**37.4%**	✗	✓	✓	−0.0021
H3+/STO-3G	Random	19	**5.89**	**9.55**	12.03	**69.0%**	**49.7%**	**36.7%**	✗	✗	✗	+0.2986
H4 chain/STO-3G	**Ours**	75	19.72	32.77	43.63	**73.7%**	**56.3%**	**41.8%**	✓	✗	✓	−0.0021
H4 chain/STO-3G	Random	75	**17.22**	**32.46**	**43.46**	**77.0%**	**56.7%**	**42.1%**	✗	✗	✗	+0.2677
LiH/STO-3G	**Ours**	179	24.38	36.02	50.75	**86.4%**	**79.9%**	**71.6%**	✓	✓	✓	+0.0012
LiH/STO-3G	Random	179	**18.33**	**33.73**	**48.31**	**89.8%**	**81.2%**	**73.0%**	✗	✗	✗	+3.0320

C.8 PREVIOUS THRESHOLD: $\tau = 0.95$

Table 19: **Matched-budget online VQE baseline comparison (previous threshold $\tau = 0.95$).** The learned selector is trained on RY/RZ VQE trajectories from depths 1–3 using row-wise top-mass target threshold $\tau = 0.95$ and evaluated in closed-loop online VQE on held-out random seeds for the same five PySCF molecules and depths. For each depth and metric, our method is paired with a Random same-budget baseline that keeps the same predicted number of QWC groups per parameter but chooses groups uniformly at random. Boldface marks the better method for each molecule-specific comparison.

Setting	Metric	Method	H2/STO-3G	HeH+/STO-3G	H3+/STO-3G	H4 chain/STO-3G	LiH/STO-3G
Depth 1	ΔE vs Full ↓	**Ours**	+0.0023	−0.0006	+0.0072	+0.0017	+0.0028
Depth 1	ΔE vs Full ↓	Random	+0.2411	+1.4286	+1.0475	+1.6624	+3.4272
Depth 1	Shot gain ↑	**Ours**	**43.8%**	**65.2%**	**74.6%**	**81.0%**	**91.5%**
Depth 1	Shot gain ↑	Random	**44.3%**	**68.5%**	**77.2%**	**85.2%**	**93.2%**
Depth 1	No-worse vs Full	**Ours**	✗	✓	✗	✗	✗
Depth 1	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 2	ΔE vs Full ↓	**Ours**	+0.0364	+0.0158	+0.0243	+0.0077	−0.0017
Depth 2	ΔE vs Full ↓	Random	+0.2499	+0.8874	+0.7010	+1.2117	+4.1073
Depth 2	Shot gain ↑	**Ours**	**40.2%**	**54.1%**	**62.0%**	**66.4%**	**88.3%**
Depth 2	Shot gain ↑	Random	**39.6%**	**51.2%**	**60.8%**	**68.8%**	**89.1%**
Depth 2	No-worse vs Full	**Ours**	✗	✗	✗	✗	✓
Depth 2	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 3	ΔE vs Full ↓	**Ours**	−0.0029	+0.0234	−0.0047	+0.0022	−0.0039
Depth 3	ΔE vs Full ↓	Random	+0.1469	+0.3339	+0.5480	+0.5354	+3.4985
Depth 3	Shot gain ↑	**Ours**	**33.6%**	**43.8%**	**50.5%**	**55.8%**	**84.1%**
Depth 3	Shot gain ↑	Random	**34.2%**	**44.1%**	**51.9%**	**58.1%**	**85.5%**
Depth 3	No-worse vs Full	**Ours**	✓	✗	✓	✗	✓
Depth 3	No-worse vs Full	Random	✗	✗	✗	✗	✗

Table 19 shows that at threshold $\tau = 0.95$, our method reaches mean absolute final-energy gap 0.00917 while saving 62.3% of full-gradient measurement shots. The matched-budget random baseline has mean absolute gap 1.33512 with 63.4% shot saving, so the selected QWC groups matter beyond the budget alone.

Table 20: **Depth-wise aggregate (previous threshold $\tau = 0.95$).** Same protocol as Table 19 with top-mass target threshold $\tau = 0.95$. Mean $|\Delta E|$ aggregates the five molecule columns at each ansatz depth; No-worse marks whether the aggregate mean signed ΔE is at most 1.6 mHa above Full.

Setting	Mean $\lvert\Delta E\rvert$ ↓	No-worse	Mean shot gain ↑
Full-QWC reference	0.00000	✓	0.0%
Ours depth 1	0.00290	✗	**71.2%**
Ours depth 2	0.01719	✗	**62.2%**
Ours depth 3	0.00742	✗	**53.6%**
Random depth 1	1.56136	✗	**73.7%**
Random depth 2	1.43145	✗	**61.9%**
Random depth 3	1.01255	✗	**54.8%**
Ours all	0.00917	✗	**62.3%**
Random all	1.33512	✗	**63.4%**

Table 21: **Detailed budget usage (previous threshold $\tau = 0.95$).** Same table format with top-mass target threshold $\tau = 0.95$. Groups/param. reports the mean selected QWC groups per parameter, shot gain is relative to full $2PGS$, and No-worse vs Full uses the one-sided 1.6 mHa criterion ($\Delta E \leq 1.6$ mHa; lower energy is accepted). Boldface marks the better Ours-vs-Random entry for each molecule-depth comparison.

Molecule	Method	Full QWC	Groups / param. ↓			Shot gain ↑			No-worse vs Full			Best ΔE ↓
		groups/param.	D1	D2	D3	D1	D2	D3	D1	D2	D3	
H2/STO-3G	**Ours**	5	2.81	**2.99**	3.32	**43.8%**	**40.2%**	**33.6%**	✗	✗	✓	+0.0023
H2/STO-3G	Random	5	**2.79**	3.02	**3.29**	**44.3%**	**39.6%**	**34.2%**	✗	✗	✗	+0.1469
HeH+/STO-3G	**Ours**	11	3.83	**5.04**	6.18	**65.2%**	**54.1%**	**43.8%**	✓	✗	✗	−0.0006
HeH+/STO-3G	Random	11	**3.47**	5.37	**6.15**	**68.5%**	**51.2%**	**44.1%**	✗	✗	✗	+0.3339
H3+/STO-3G	**Ours**	19	4.82	**7.23**	9.41	**74.6%**	**62.0%**	**50.5%**	✗	✗	✓	−0.0047
H3+/STO-3G	Random	19	**4.34**	7.44	**9.13**	**77.2%**	**60.8%**	**51.9%**	✗	✗	✗	+0.5480
H4 chain/STO-3G	**Ours**	75	14.28	25.16	33.14	**81.0%**	**66.4%**	**55.8%**	✗	✗	✗	+0.0017
H4 chain/STO-3G	Random	75	**11.08**	**23.37**	**31.45**	**85.2%**	**68.8%**	**58.1%**	✗	✗	✗	+0.5354
LiH/STO-3G	**Ours**	179	15.21	20.93	28.47	**91.5%**	**88.3%**	**84.1%**	✗	✓	✓	−0.0017
LiH/STO-3G	Random	179	**12.21**	**19.58**	**25.95**	**93.2%**	**89.1%**	**85.5%**	✗	✗	✗	+3.4272

C.9 PREVIOUS THRESHOLD: $\tau = 0.99$

Table 22: **Matched-budget online VQE baseline comparison (conservative threshold $\tau = 0.99$).** The learned selector is trained on RY/RZ VQE trajectories from depths 1–3 using row-wise top-mass target threshold $\tau = 0.99$ and evaluated in closed-loop online VQE on held-out random seeds for the same five PySCF molecules and depths. For each depth and metric, our method is paired with a Random same-budget baseline that keeps the same predicted number of QWC groups per parameter but chooses groups uniformly at random. Boldface marks the better method for each molecule-specific comparison.

Setting	Metric	Method	H2/STO-3G	HeH+/STO-3G	H3+/STO-3G	H4 chain/STO-3G	LiH/STO-3G
Depth 1	ΔE vs Full ↓	**Ours**	+0.0013	+0.0013	+0.0008	−0.0008	+0.0013
Depth 1	ΔE vs Full ↓	Random	+0.2060	+1.0934	+0.5968	+1.0145	+3.0269
Depth 1	Shot gain ↑	**Ours**	**36.6%**	**51.3%**	**58.2%**	**67.4%**	**81.7%**
Depth 1	Shot gain ↑	Random	**35.5%**	**54.3%**	**60.1%**	**72.7%**	**86.9%**
Depth 1	No-worse vs Full	**Ours**	✓	✓	✓	✓	✓
Depth 1	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 2	ΔE vs Full ↓	**Ours**	−0.0008	+0.0037	−0.0015	+0.0012	−0.0009
Depth 2	ΔE vs Full ↓	Random	+0.1683	+0.3132	+0.3132	+0.5964	+3.4087
Depth 2	Shot gain ↑	**Ours**	**25.8%**	**32.1%**	**41.2%**	**47.7%**	**74.7%**
Depth 2	Shot gain ↑	Random	**25.2%**	**32.3%**	**41.6%**	**47.2%**	**75.7%**
Depth 2	No-worse vs Full	**Ours**	✓	✗	✓	✓	✓
Depth 2	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 3	ΔE vs Full ↓	**Ours**	+0.0028	+0.0087	−0.0006	−0.0009	+0.0008
Depth 3	ΔE vs Full ↓	Random	+0.1022	+0.2128	+0.1860	+0.1159	+2.6667
Depth 3	Shot gain ↑	**Ours**	**18.0%**	**23.6%**	**28.5%**	**33.1%**	**63.1%**
Depth 3	Shot gain ↑	Random	**17.4%**	**23.4%**	**28.2%**	**33.0%**	**64.7%**
Depth 3	No-worse vs Full	**Ours**	✗	✗	✓	✓	✓
Depth 3	No-worse vs Full	Random	✗	✗	✗	✗	✗

Table 22 shows that at threshold $\tau = 0.99$, our method reaches mean absolute final-energy gap 0.00439 while saving 45.5% of full-gradient measurement shots. The matched-budget random baseline has mean absolute gap 0.93589 with 46.5% shot saving, so the selected QWC groups matter beyond the budget alone.

Table 23: **Depth-wise aggregate (conservative threshold $\tau = 0.99$).** Same protocol as Table 22 with top-mass target threshold $\tau = 0.99$. Mean $|\Delta E|$ aggregates the five molecule columns at each ansatz depth; No-worse marks whether the aggregate mean signed ΔE is at most 1.6 mHa above Full.

Setting	Mean $\lvert\Delta E\rvert$ ↓	No-worse	Mean shot gain ↑
Full-QWC reference	0.00000	✓	0.0%
Ours depth 1	0.00139	✓	**59.0%**
Ours depth 2	0.00560	✓	**44.3%**
Ours depth 3	0.00620	✗	**33.2%**
Random depth 1	1.18750	✗	**61.9%**
Random depth 2	0.95994	✗	**44.4%**
Random depth 3	0.66023	✗	**33.3%**
Ours all	0.00439	✓	**45.5%**
Random all	0.93589	✗	**46.5%**

Table 24: **Detailed budget usage (conservative threshold $\tau = 0.99$).** Same table format with top-mass target threshold $\tau = 0.99$. Groups/param. reports the mean selected QWC groups per parameter, shot gain is relative to full $2PGS$, and No-worse vs Full uses the one-sided 1.6 mHa criterion ($\Delta E \leq 1.6$ mHa; lower energy is accepted). Boldface marks the better Ours-vs-Random entry for each molecule-depth comparison.

Molecule	Method	Full QWC	Groups / param. ↓			Shot gain ↑			No-worse vs Full			Best ΔE ↓
		groups/param.	D1	D2	D3	D1	D2	D3	D1	D2	D3	
H2/STO-3G	**Ours**	5	**3.17**	**3.71**	**4.10**	**36.6%**	**25.8%**	**18.0%**	✓	✓	✗	+0.0013
H2/STO-3G	Random	5	3.22	3.74	4.13	**35.5%**	**25.2%**	**17.4%**	✗	✗	✗	+0.1022
HeH+/STO-3G	**Ours**	11	5.35	7.47	**8.41**	**51.3%**	**32.1%**	**23.6%**	✓	✗	✗	+0.0013
HeH+/STO-3G	Random	11	**5.03**	**7.45**	8.42	**54.3%**	**32.3%**	**23.4%**	✗	✗	✗	+0.2128
H3+/STO-3G	**Ours**	19	7.94	11.16	**13.59**	**58.2%**	**41.2%**	**28.5%**	✓	✓	✓	+0.0008
H3+/STO-3G	Random	19	**7.58**	**11.09**	13.65	**60.1%**	**41.6%**	**28.2%**	✗	✗	✗	+0.1860
H4 chain/STO-3G	**Ours**	75	24.45	**39.22**	**50.17**	**67.4%**	**47.7%**	**33.1%**	✓	✓	✓	−0.0008
H4 chain/STO-3G	Random	75	**20.45**	39.59	50.24	**72.7%**	**47.2%**	**33.0%**	✗	✗	✗	+0.1159
LiH/STO-3G	**Ours**	179	32.84	45.34	66.09	**81.7%**	**74.7%**	**63.1%**	✓	✓	✓	+0.0013
LiH/STO-3G	Random	179	**23.47**	**43.46**	**63.21**	**86.9%**	**75.7%**	**64.7%**	✗	✗	✗	+2.6667

Table 25: **Molecule-wise shot-allocation benchmark at $b = 1.0$.** This expands Table 2 using the layout of Table 1: five STO-3G molecules, depths 1–2, four held-out seeds, 28 online iterations, and 1024 nominal shots per shifted selected parameter–group setting. ΔE and shot gain are paired against Full-QWC uniform; adaptive allocators choose their realized shot counts. ✓/✗ applies the one-sided mean criterion $\Delta E \leq 1.6$ mHa. Boldface marks the closest non-Full mean energy and bold-underlining marks the largest shot gain in each molecule–depth cell.

Setting	Metric	Method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 1	ΔE vs Full ↓	Full-QWC uniform	+0.0000	+0.0000	+0.0000	+0.0000	+0.0000
Depth 1	ΔE vs Full ↓	iCANS (Kübler et al.)	+0.0038	+0.0793	+0.0321	+0.0086	+0.0177
Depth 1	ΔE vs Full ↓	gCANS (Gu et al.)	+0.0205	+0.0881	+0.0703	−0.0062	+0.0315
Depth 1	ΔE vs Full ↓	Rosalin (Arrasmith et al.)	+0.0070	−0.0092	+0.0005	+0.0060	+0.0076
Depth 1	ΔE vs Full ↓	Variance-aware (Zhu et al.)	−0.0028	−0.0116	+0.0020	+0.0033	+0.0136
Depth 1	ΔE vs Full ↓	**Ours** uniform	−0.0024	−0.0221	−0.0157	−0.0082	+0.0006
Depth 1	ΔE vs Full ↓	**Ours**+variance-aware (Zhu et al.)	−0.0034	−0.0232	−0.0230	−0.0137	+0.0024
Depth 1	ΔE vs Full ↓	**Ours**+gCANS (Gu et al.)	+0.0108	+0.0335	+0.0145	+0.0099	−0.0020
Depth 1	Shot gain ↑	Full-QWC uniform	0.0%	0.0%	0.0%	0.0%	0.0%
Depth 1	Shot gain ↑	iCANS (Kübler et al.)	−73.0%	−64.3%	−73.8%	−89.1%	−108.9%
Depth 1	Shot gain ↑	gCANS (Gu et al.)	92.7%	93.8%	93.3%	93.0%	90.2%
Depth 1	Shot gain ↑	Rosalin (Arrasmith et al.)	−85.0%	−62.8%	−90.3%	−114.7%	−135.7%
Depth 1	Shot gain ↑	Variance-aware (Zhu et al.)	0.0%	0.0%	0.0%	0.0%	0.0%
Depth 1	Shot gain ↑	**Ours** uniform	41.2%	59.6%	64.2%	76.2%	87.9%
Depth 1	Shot gain ↑	**Ours**+variance-aware (Zhu et al.)	41.5%	59.7%	63.7%	75.8%	87.7%
Depth 1	Shot gain ↑	**Ours**+gCANS (Gu et al.)	**96.3%**	**97.6%**	**97.8%**	**98.5%**	**99.3%**
Depth 1	No-worse vs Full	Full-QWC uniform	✓	✓	✓	✓	✓
Depth 1	No-worse vs Full	iCANS (Kübler et al.)	✗	✗	✗	✗	✗
Depth 1	No-worse vs Full	gCANS (Gu et al.)	✗	✗	✗	✓	✗
Depth 1	No-worse vs Full	Rosalin (Arrasmith et al.)	✗	✓	✓	✗	✗
Depth 1	No-worse vs Full	Variance-aware (Zhu et al.)	✓	✓	✗	✗	✗
Depth 1	No-worse vs Full	**Ours** uniform	✓	✓	✓	✓	✓
Depth 1	No-worse vs Full	**Ours**+variance-aware (Zhu et al.)	✓	✓	✓	✓	✗
Depth 1	No-worse vs Full	**Ours**+gCANS (Gu et al.)	✗	✗	✗	✗	✓
Depth 2	ΔE vs Full ↓	Full-QWC uniform	+0.0000	+0.0000	+0.0000	+0.0000	+0.0000
Depth 2	ΔE vs Full ↓	iCANS (Kübler et al.)	+0.1044	+0.1569	+0.0239	+0.1216	+0.2486
Depth 2	ΔE vs Full ↓	gCANS (Gu et al.)	+0.0279	−0.0474	+0.0853	+0.1612	+0.3004
Depth 2	ΔE vs Full ↓	Rosalin (Arrasmith et al.)	+0.0570	+0.1309	−0.0221	+0.0157	−0.1710
Depth 2	ΔE vs Full ↓	Variance-aware (Zhu et al.)	+0.0645	+0.0997	+0.0050	+0.0044	−0.1812
Depth 2	ΔE vs Full ↓	**Ours** uniform	+0.0954	+0.0222	−0.0962	+0.0720	+0.0427
Depth 2	ΔE vs Full ↓	**Ours**+variance-aware (Zhu et al.)	+0.0336	+0.0237	−0.0175	−0.0000	−0.1850
Depth 2	ΔE vs Full ↓	**Ours**+gCANS (Gu et al.)	−0.0319	+0.0938	+0.0138	+0.1896	+0.0891
Depth 2	Shot gain ↑	Full-QWC uniform	0.0%	0.0%	0.0%	0.0%	0.0%
Depth 2	Shot gain ↑	iCANS (Kübler et al.)	−67.5%	−13.3%	−60.1%	−71.6%	−144.6%
Depth 2	Shot gain ↑	gCANS (Gu et al.)	92.8%	93.7%	93.1%	90.4%	84.6%
Depth 2	Shot gain ↑	Rosalin (Arrasmith et al.)	−68.1%	−42.6%	−86.8%	−110.4%	−181.2%
Depth 2	Shot gain ↑	Variance-aware (Zhu et al.)	0.0%	0.0%	0.0%	0.0%	0.0%
Depth 2	Shot gain ↑	**Ours** uniform	27.8%	33.2%	46.3%	52.0%	82.4%
Depth 2	Shot gain ↑	**Ours**+variance-aware (Zhu et al.)	27.5%	31.9%	45.6%	52.1%	82.8%
Depth 2	Shot gain ↑	**Ours**+gCANS (Gu et al.)	**95.3%**	**95.8%**	**96.6%**	**96.4%**	**98.3%**
Depth 2	No-worse vs Full	Full-QWC uniform	✓	✓	✓	✓	✓
Depth 2	No-worse vs Full	iCANS (Kübler et al.)	✗	✗	✗	✗	✗
Depth 2	No-worse vs Full	gCANS (Gu et al.)	✗	✓	✗	✗	✗
Depth 2	No-worse vs Full	Rosalin (Arrasmith et al.)	✗	✗	✓	✗	✓
Depth 2	No-worse vs Full	Variance-aware (Zhu et al.)	✗	✗	✗	✗	✓
Depth 2	No-worse vs Full	**Ours** uniform	✗	✗	✓	✗	✗
Depth 2	No-worse vs Full	**Ours**+variance-aware (Zhu et al.)	✗	✗	✓	✓	✓
Depth 2	No-worse vs Full	**Ours**+gCANS (Gu et al.)	✓	✗	✗	✗	✗

D Shot-Allocation Details

D.1 Molecule-Wise In-Distribution Allocation

D.2 Depth-OOD Shot-Allocation Benchmark

Table 26: **Shot-allocation benchmark in the Table 29 depth-OOD protocol at $b = 1.0$.** The saved selector is trained on depths 1–4 and evaluated on unseen depths 5 and 6 over five PySCF/STO-3G molecules and four held-out seeds; depth 7 is reported in Appendix D.3. All methods use 28 common Adam updates and the same moment-matched Gaussian finite-shot estimator with 1024 nominal shots per selected parameter–group pair and $\tau = 0.997$. iCANS (Kübler et al., 2020), gCANS (Gu et al., 2021), Rosalin (Arrasmith et al., 2020), and variance-aware allocation (Zhu et al., 2024) are controlled allocation policies rather than reproductions of their complete optimizer stacks. Green/red ΔE and ✓/✗ use the one-sided 1.6 mHa criterion. Bold marks the closest non-Full mean ΔE; bold-underlining marks the largest saving in each cell.

Setting	Metric	Method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 5	ΔE vs Full ↓	Full-QWC uniform	+0.0000	+0.0000	+0.0000	+0.0000	+0.0000
Depth 5	ΔE vs Full ↓	iCANS	+0.0380	+0.0300	+0.0334	−0.0279	−0.0371
Depth 5	ΔE vs Full ↓	gCANS	+0.0690	+0.1606	+0.0775	+0.1232	+0.0538
Depth 5	ΔE vs Full ↓	Rosalin	−0.0192	−0.0242	−0.0014	−0.0212	−0.2266
Depth 5	ΔE vs Full ↓	Variance-aware	−0.0232	−0.0382	−0.0447	−0.1764	−0.2449
Depth 5	ΔE vs Full ↓	**Ours** uniform	−0.0026	+0.0484	−0.0274	−0.0407	−0.0139
Depth 5	ΔE vs Full ↓	**Ours**+variance-aware	−0.0094	−0.0339	−0.0543	−0.1588	−0.2644
Depth 5	ΔE vs Full ↓	**Ours**+gCANS	+0.0158	+0.0569	−0.0597	−0.1106	−0.2085
Depth 5	Shot gain ↑	Full-QWC uniform	0.0%	0.0%	0.0%	0.0%	0.0%
Depth 5	Shot gain ↑	iCANS	−37.2%	−8.0%	−79.9%	−116.5%	−190.5%
Depth 5	Shot gain ↑	gCANS	89.9%	93.2%	73.7%	57.1%	39.5%
Depth 5	Shot gain ↑	Rosalin	−43.2%	−21.3%	−104.4%	−162.1%	−229.2%
Depth 5	Shot gain ↑	Variance-aware	0.0%	0.0%	0.0%	0.0%	0.0%
Depth 5	Shot gain ↑	**Ours** uniform	5.6%	8.5%	10.6%	17.1%	55.7%
Depth 5	Shot gain ↑	**Ours**+variance-aware	5.6%	8.5%	10.6%	17.0%	53.7%
Depth 5	Shot gain ↑	**Ours**+gCANS	**90.2%**	**93.6%**	**78.7%**	**61.6%**	**68.8%**
Depth 5	No-worse vs Full	Full-QWC uniform	✓	✓	✓	✓	✓
Depth 5	No-worse vs Full	iCANS	✗	✗	✗	✓	✓
Depth 5	No-worse vs Full	gCANS	✗	✗	✗	✗	✗
Depth 5	No-worse vs Full	Rosalin	✓	✓	✓	✓	✓
Depth 5	No-worse vs Full	Variance-aware	✓	✓	✓	✓	✓
Depth 5	No-worse vs Full	**Ours** uniform	✓	✗	✓	✓	✓
Depth 5	No-worse vs Full	**Ours**+variance-aware	✓	✓	✓	✓	✓
Depth 5	No-worse vs Full	**Ours**+gCANS	✗	✗	✓	✓	✓
Depth 6	ΔE vs Full ↓	Full-QWC uniform	+0.0000	+0.0000	+0.0000	+0.0000	+0.0000
Depth 6	ΔE vs Full ↓	iCANS	+0.0659	+0.0090	+0.0690	+0.1038	+0.0150
Depth 6	ΔE vs Full ↓	gCANS	+0.0555	+0.0643	+0.1207	+0.0554	+0.2048
Depth 6	ΔE vs Full ↓	Rosalin	+0.0101	−0.0008	+0.0200	−0.0183	−0.3586
Depth 6	ΔE vs Full ↓	Variance-aware	+0.0021	−0.0258	−0.0009	+0.0039	−0.2956
Depth 6	ΔE vs Full ↓	**Ours** uniform	+0.0320	−0.0262	+0.0307	+0.0214	+0.0280
Depth 6	ΔE vs Full ↓	**Ours**+variance-aware	+0.0084	−0.0126	−0.0173	−0.0265	−0.3267
Depth 6	ΔE vs Full ↓	**Ours**+gCANS	+0.0385	−0.0262	−0.0305	−0.0211	−0.2811
Depth 6	Shot gain ↑	Full-QWC uniform	0.0%	0.0%	0.0%	0.0%	0.0%
Depth 6	Shot gain ↑	iCANS	−58.3%	−7.4%	−97.1%	−133.9%	−196.7%
Depth 6	Shot gain ↑	gCANS	89.7%	92.8%	57.3%	38.0%	8.3%
Depth 6	Shot gain ↑	Rosalin	−59.3%	−29.8%	−110.5%	−181.7%	−241.1%
Depth 6	Shot gain ↑	Variance-aware	0.0%	0.0%	0.0%	0.0%	0.0%
Depth 6	Shot gain ↑	**Ours** uniform	4.8%	7.6%	8.7%	14.5%	**46.1%**
Depth 6	Shot gain ↑	**Ours**+variance-aware	4.8%	7.6%	8.7%	14.5%	45.1%
Depth 6	Shot gain ↑	**Ours**+gCANS	**89.9%**	**93.1%**	**61.6%**	**41.4%**	36.4%
Depth 6	No-worse vs Full	Full-QWC uniform	✓	✓	✓	✓	✓
Depth 6	No-worse vs Full	iCANS	✗	✗	✗	✗	✗
Depth 6	No-worse vs Full	gCANS	✗	✗	✗	✗	✗
Depth 6	No-worse vs Full	Rosalin	✗	✓	✗	✓	✓
Depth 6	No-worse vs Full	Variance-aware	✗	✓	✓	✗	✓
Depth 6	No-worse vs Full	**Ours** uniform	✗	✓	✗	✗	✗
Depth 6	No-worse vs Full	**Ours**+variance-aware	✗	✓	✓	✓	✓
Depth 6	No-worse vs Full	**Ours**+gCANS	✗	✓	✓	✓	✓

Across all 60 held-out seed-level cases, **Ours** uniform is the most faithful non-Full policy, with mean $|\Delta E| = 0.0568$ Ha while saving 16.6% of shots. Composing selection with gCANS increases saving to 63.3% and improves on gCANS alone in mean $|\Delta E|$ (0.0973 vs. 0.1014 Ha) and seed-level no-worse rate (60.0% vs. 10.0%). The variance-aware composition reaches the highest Ours-family no-worse rate (68.3%) but with a larger absolute gap. Thus Ours-uniform is the clean fidelity–saving point, whereas Ours+gCANS is the aggressive Pareto point.

D.3 Depth-7 Shot-Allocation Detail

Table 27: **Depth-7 continuation of the depth-OOD shot-allocation benchmark.** This table uses the same trained depth-1–4 selector, five PySCF/STO-3G molecules, four held-out seeds, finite-shot estimator, $b = 1.0$, and $\tau = 0.997$ as Table 26.

Setting	Metric	Method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 7	ΔE vs Full ↓	Full-QWC uniform	+0.0000	+0.0000	+0.0000	+0.0000	+0.0000
Depth 7	ΔE vs Full ↓	iCANS	+0.1195	+0.0256	+0.0540	+0.0295	+0.0791
Depth 7	ΔE vs Full ↓	gCANS	+0.0151	+0.1313	+0.0640	+0.0598	+0.1809
Depth 7	ΔE vs Full ↓	Rosalin	+0.0517	+0.0421	−0.0208	−0.0153	−0.2483
Depth 7	ΔE vs Full ↓	Variance-aware	−0.0276	−0.0240	−0.0199	−0.0429	−0.2975
Depth 7	ΔE vs Full ↓	**Ours** uniform	−0.0109	−0.0544	+0.0231	+0.0099	+0.0250
Depth 7	ΔE vs Full ↓	**Ours**+variance-aware	−0.0125	−0.0392	−0.0045	−0.0712	−0.2989
Depth 7	ΔE vs Full ↓	**Ours**+gCANS	−0.0025	+0.0291	+0.0091	−0.0303	−0.3097
Depth 7	Shot gain ↑	Full-QWC uniform	0.0%	0.0%	0.0%	0.0%	0.0%
Depth 7	Shot gain ↑	iCANS	−48.9%	−12.8%	−89.5%	−147.0%	−200.9%
Depth 7	Shot gain ↑	gCANS	**87.2%**	**92.0%**	**65.0%**	5.3%	−31.4%
Depth 7	Shot gain ↑	Rosalin	−52.0%	−31.4%	−111.2%	−191.9%	−249.3%
Depth 7	Shot gain ↑	Variance-aware	0.0%	0.0%	0.0%	0.0%	0.0%
Depth 7	Shot gain ↑	**Ours** uniform	3.8%	6.7%	7.5%	13.1%	**39.2%**
Depth 7	Shot gain ↑	**Ours**+variance-aware	3.6%	6.6%	7.4%	**13.1%**	38.0%
Depth 7	Shot gain ↑	**Ours**+gCANS	86.5%	91.6%	64.0%	−0.0%	−8.2%
Depth 7	No-worse vs Full	Full-QWC uniform	✓	✓	✓	✓	✓
Depth 7	No-worse vs Full	iCANS	✗	✗	✗	✗	✗
Depth 7	No-worse vs Full	gCANS	✗	✗	✗	✗	✗
Depth 7	No-worse vs Full	Rosalin	✗	✗	✓	✓	✓
Depth 7	No-worse vs Full	Variance-aware	✓	✓	✓	✓	✓
Depth 7	No-worse vs Full	**Ours** uniform	✓	✓	✗	✗	✗
Depth 7	No-worse vs Full	**Ours**+variance-aware	✓	✓	✓	✓	✓
Depth 7	No-worse vs Full	**Ours**+gCANS	✓	✗	✗	✓	✓

Table 28: **Shot-allocation benchmark on the Table 39 Hamiltonian panel at $b = 1.0$.** We reuse the saved depth-1–4 selector and evaluate seven held-out 8-qubit Hamiltonians at depths 1 and 4 over seeds 97/101/107/109, with 28 Adam updates, $\tau = 0.997$, and 1024 nominal shots per shifted selected parameter–group setting. All methods are rerun against Full-QWC uniform with the same moment-matched Gaussian finite-shot estimator and common Adam optimizer; iCANS (Kübler et al., 2020), gCANS (Gu et al., 2021), Rosalin (Arrasmith et al., 2020), and variance-aware allocation (Zhu et al., 2024) are therefore controlled allocation policies rather than reproductions of their complete optimizer stacks. Green/red ΔE and ✓/✗ use the one-sided 1.6 mHa criterion. Bold marks the closest non-Full mean ΔE; bold-underlining marks the largest shot gain in each cell.

Testing Hamiltonian	Depth	Metric	Full-U	iCANS	gCANS	Rosalin	Var.-aware	**Ours**-U	**Ours**+Var.	**Ours**+gCANS
TFIM $h = 0.5$	Depth 1	ΔE vs Full ↓	+0.0000	−0.0120	−0.0000	−0.0050	+0.0041	−0.0013	−0.0012	+0.0009
	Depth 1	Shot gain ↑	0.0%	−126.6%	**72.1%**	−106.4%	0.0%	2.1%	2.0%	69.9%
	Depth 1	No-worse vs Full	✓	✓	✓	✓	✗	✓	✓	✓
	Depth 4	ΔE vs Full ↓	+0.0000	+0.0245	+0.0427	+0.0028	−0.0071	−0.0002	−0.0022	+0.0445
	Depth 4	Shot gain ↑	0.0%	−88.5%	77.0%	−95.0%	0.0%	2.2%	1.9%	**77.4%**
	Depth 4	No-worse vs Full	✓	✗	✗	✗	✓	✓	✓	✗
TFIM $h = 0.7$	Depth 1	ΔE vs Full ↓	+0.0000	−0.0122	+0.0081	−0.0031	+0.0093	−0.0020	+0.0014	+0.0192
	Depth 1	Shot gain ↑	0.0%	−116.5%	**81.4%**	−91.5%	0.0%	3.2%	3.1%	80.9%
	Depth 1	No-worse vs Full	✓	✓	✗	✓	✗	✓	✓	✗
	Depth 4	ΔE vs Full ↓	+0.0000	+0.0182	+0.0718	+0.0087	−0.0129	−0.0161	−0.0179	+0.0606
	Depth 4	Shot gain ↑	0.0%	−88.5%	73.2%	−81.2%	0.0%	2.1%	1.9%	**75.6%**
	Depth 4	No-worse vs Full	✓	✗	✗	✗	✓	✓	✓	✗
TFIM $h = 0.9$	Depth 1	ΔE vs Full ↓	+0.0000	+0.0041	+0.0300	−0.0037	+0.0163	+0.0811	+0.2016	+0.1387
	Depth 1	Shot gain ↑	0.0%	−106.5%	82.7%	−80.4%	0.0%	25.0%	25.0%	**88.7%**
	Depth 1	No-worse vs Full	✓	✗	✗	✓	✗	✗	✗	✗
	Depth 4	ΔE vs Full ↓	+0.0000	+0.0321	+0.0813	+0.0561	−0.0146	+0.0496	+0.0307	+0.0691
	Depth 4	Shot gain ↑	0.0%	−88.3%	72.3%	−76.2%	0.0%	6.2%	6.2%	**79.5%**
	Depth 4	No-worse vs Full	✓	✗	✗	✗	✓	✗	✗	✗
Heis. XXZ+X field	Depth 1	ΔE vs Full ↓	+0.0000	+0.4624	+0.7663	+0.3730	−0.1707	−0.2163	+0.1527	+0.6480
	Depth 1	Shot gain ↑	0.0%	−55.0%	89.8%	−62.1%	0.0%	16.7%	16.7%	**92.8%**
	Depth 1	No-worse vs Full	✓	✗	✗	✗	✓	✓	✗	✗
	Depth 4	ΔE vs Full ↓	+0.0000	+0.5043	+1.1720	−0.0388	+0.0759	−0.0737	−0.0484	+0.8442
	Depth 4	Shot gain ↑	0.0%	−59.6%	82.2%	−61.5%	0.0%	4.2%	4.2%	**82.2%**
	Depth 4	No-worse vs Full	✓	✗	✗	✓	✗	✓	✓	✗
Heis. J1–J2	Depth 1	ΔE vs Full ↓	+0.0000	+0.2970	+0.6094	−0.0259	−0.3164	−0.1839	−0.0655	+0.3957
	Depth 1	Shot gain ↑	0.0%	−49.9%	91.4%	−53.7%	0.0%	16.7%	16.7%	**93.9%**
	Depth 1	No-worse vs Full	✓	✗	✗	✓	✓	✓	✓	✗
	Depth 4	ΔE vs Full ↓	+0.0000	+0.1698	+1.0128	+0.3410	+0.2490	+0.1508	+0.3493	+0.8673
	Depth 4	Shot gain ↑	0.0%	−52.0%	80.7%	−64.7%	0.0%	4.2%	4.2%	**80.9%**
	Depth 4	No-worse vs Full	✓	✗	✗	✗	✗	✗	✗	✗
Heis. XXY+Z field	Depth 1	ΔE vs Full ↓	+0.0000	+0.3060	+0.2416	+0.2757	−0.2531	−0.0757	+0.0825	+0.5081
	Depth 1	Shot gain ↑	0.0%	−69.4%	86.0%	−63.3%	0.0%	16.7%	16.7%	**89.5%**
	Depth 1	No-worse vs Full	✓	✗	✗	✗	✓	✓	✗	✗
	Depth 4	ΔE vs Full ↓	+0.0000	+0.3343	+0.4644	+0.1376	+0.3029	−0.4193	−0.1713	+0.8496
	Depth 4	Shot gain ↑	0.0%	−69.7%	**77.0%**	−59.8%	0.0%	4.2%	4.2%	74.5%
	Depth 4	No-worse vs Full	✓	✗	✗	✗	✗	✓	✓	✗
JW-like parity spin	Depth 1	ΔE vs Full ↓	+0.0000	+0.0465	+0.1333	+0.0172	−0.0230	+0.0185	−0.0198	+0.0440
	Depth 1	Shot gain ↑	0.0%	−94.4%	92.7%	−95.6%	0.0%	36.2%	36.1%	**95.7%**
	Depth 1	No-worse vs Full	✓	✗	✗	✗	✓	✗	✓	✗
	Depth 4	ΔE vs Full ↓	+0.0000	−0.0141	+0.1421	−0.0724	−0.0814	−0.1151	−0.0997	+0.0190
	Depth 4	Shot gain ↑	0.0%	−104.6%	72.2%	−118.7%	0.0%	11.5%	11.6%	**75.2%**
	Depth 4	No-worse vs Full	✓	✓	✗	✓	✓	✓	✓	✗

D.4 Hamiltonian-Panel Allocation

Across all 56 paired seed-level cases, **Ours**-uniform saves 10.8% of shots with mean $|\Delta E| = 0.1889$ Ha, while Full-QWC variance-aware allocation has a smaller 0.1614 Ha gap but no shot saving. Composing our method with gCANS reaches an aggressive 82.6% saving and slightly improves on gCANS alone in both saving (80.8%) and mean absolute gap (0.3748 vs. 0.4001 Ha), but both pass only 14.3% of seed-level comparisons. In contrast, iCANS and Rosalin spend 83.5% and 79.3% more shots than Full on average at this operating point. The panel therefore supports composability, while also showing that allocation policies calibrated on easier molecular regimes do not automatically preserve energy under Hamiltonian-family shift.

E DEPTH-OOD TRANSFER

E.1 DEPTH-OOD TRANSFER BASELINES

This depth-OOD benchmark evaluates the conservative $\tau = 0.997$ selector on unseen RY/RZ depths 5–7; the table reports depths 5 and 6. A matched-budget control adds a cell-wise cost-matched TITAN comparison: TITAN† chooses the freezing threshold whose shot gain is closest to **Ours**, while TITAN†+**Ours** uses that same active-parameter mask and additionally sparsifies QWC groups.

Table 29: **Depth-OOD online VQE transfer baselines at $\tau = 0.997$.** The relevance model is trained on depths 1–4; the table reports unseen depths 5 and 6 over five PySCF/STO-3G molecules and four held-out seeds. Random is group-budget matched to our method. TITAN† uses a cell-wise freezing threshold selected to match our method's shot gain; TITAN†+**Ours** reuses that threshold and adds group selection. Green/red ΔE indicates the one-sided 1.6 mHa no-worse criterion. Bold marks the closest mean ΔE to Full; bold-underlined shot gain marks the largest saving.

Setting	Metric	Method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 5	ΔE vs Full ↓	**Ours**	**-0.0003**	+0.0053	**-0.0120**	**+0.0012**	**-0.0068**
Depth 5	ΔE vs Full ↓	Random	+0.0179	+0.0684	+0.0503	+0.1580	+2.1893
Depth 5	ΔE vs Full ↓	TITAN (Peng et al.)	+0.0239	+0.0130	+0.0146	+0.1011	+0.2254
Depth 5	ΔE vs Full ↓	TITAN+**Ours**	+0.0251	**+0.0048**	+0.0111	+0.1009	+0.2250
Depth 5	ΔE vs Full ↓	TITAN†	**-0.0028**	**-0.0059**	**-0.0044**	+0.0416	+0.0914
Depth 5	ΔE vs Full ↓	TITAN†+**Ours**	**-0.0035**	**-0.0056**	**-0.0112**	+0.0392	+0.0928
Depth 5	Shot gain ↑	**Ours**	5.7%	8.4%	10.6%	17.1%	52.9%
Depth 5	Shot gain ↑	Random	5.7%	8.5%	10.7%	17.2%	55.4%
Depth 5	Shot gain ↑	TITAN (Peng et al.)	32.0%	32.7%	38.7%	42.3%	73.4%
Depth 5	Shot gain ↑	TITAN+**Ours**	**34.3%**	**37.9%**	**44.9%**	**52.2%**	**93.2%**
Depth 5	Shot gain ↑	TITAN†	6.9%	12.7%	13.1%	14.4%	52.4%
Depth 5	Shot gain ↑	TITAN†+**Ours**	10.3%	19.5%	22.2%	28.4%	83.2%
Depth 5	No-worse vs Full	**Ours**	✓	✗	✓	✓	✓
Depth 5	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 5	No-worse vs Full	TITAN (Peng et al.)	✗	✗	✗	✗	✗
Depth 5	No-worse vs Full	TITAN+**Ours**	✗	✗	✗	✗	✗
Depth 5	No-worse vs Full	TITAN†	✓	✓	✓	✗	✗
Depth 5	No-worse vs Full	TITAN†+**Ours**	✓	✓	✓	✗	✗
Depth 6	ΔE vs Full ↓	**Ours**	**+0.0000**	**-0.0036**	**+0.0003**	**-0.0081**	**+0.0086**
Depth 6	ΔE vs Full ↓	Random	+0.0046	**+0.0005**	+0.0048	+0.0785	+0.7440
Depth 6	ΔE vs Full ↓	TITAN (Peng et al.)	+0.0139	**-0.0088**	**-0.0426**	+0.0070	+0.3083
Depth 6	ΔE vs Full ↓	TITAN+**Ours**	+0.0133	**-0.0113**	**-0.0429**	+0.0082	+0.3092
Depth 6	ΔE vs Full ↓	TITAN†	**-0.0070**	**-0.0242**	+0.0130	**-0.0034**	+0.0825
Depth 6	ΔE vs Full ↓	TITAN†+**Ours**	**-0.0069**	**-0.0280**	+0.0162	**-0.0071**	+0.0798
Depth 6	Shot gain ↑	**Ours**	4.8%	7.6%	8.7%	14.6%	44.3%
Depth 6	Shot gain ↑	Random	4.7%	7.5%	8.7%	14.6%	46.3%
Depth 6	Shot gain ↑	TITAN (Peng et al.)	32.8%	30.5%	28.1%	39.4%	71.2%
Depth 6	Shot gain ↑	TITAN+**Ours**	**34.6%**	**34.4%**	**33.8%**	**47.6%**	**88.5%**
Depth 6	Shot gain ↑	TITAN†	7.0%	9.6%	7.5%	12.4%	37.0%
Depth 6	Shot gain ↑	TITAN†+**Ours**	10.1%	15.5%	15.0%	24.8%	67.6%
Depth 6	No-worse vs Full	**Ours**	✓	✓	✓	✓	✗
Depth 6	No-worse vs Full	Random	✗	✓	✗	✗	✗
Depth 6	No-worse vs Full	TITAN (Peng et al.)	✗	✓	✓	✗	✗
Depth 6	No-worse vs Full	TITAN+**Ours**	✗	✓	✓	✗	✗
Depth 6	No-worse vs Full	TITAN†	✓	✓	✗	✓	✗
Depth 6	No-worse vs Full	TITAN†+**Ours**	✓	✓	✗	✓	✗

Table 29 separates parameter-specific QWC selection from parameter freezing at both the original and cost-matched operating points. The dagger rows remove shot budget as an explanation for differences between our method and TITAN-only training; the composed row shows the additional saving available when the two orthogonal controls are combined.

E.2 Additional Depth-OOD Transfer Tables

The tables below report aggregate performance and budget usage for depth-OOD transfer at $\tau = 0.995$.

Table 30: **Depth-OOD aggregate at $\tau = 0.995$.** Mean $|\Delta E|$ averages the five molecule columns at each unseen depth. No-worse uses the one-sided criterion $\Delta E \leq 1.6$ mHa against the Full-QWC reference, so lower final energy is accepted.

Setting	Mean $\lvert\Delta E\rvert \downarrow$	No-worse	Mean shot gain ↑
Full-QWC reference	0.00000	✓	0.0%
Ours depth 5	0.01540	✓	**22.6%**
Ours depth 6	0.00697	✓	**19.6%**
Ours depth 7	0.00684	✓	**17.3%**
Random depth 5	0.54065	✗	**23.5%**
Random depth 6	0.26326	✗	**20.1%**
Random depth 7	0.15211	✗	**17.7%**
Ours all	0.00974	✓	**19.8%**
Random all	0.31867	✗	**20.4%**

Table 31: **Depth-OOD budget usage at $\tau = 0.995$.** Groups/param. reports the mean selected QWC groups per parameter on unseen depths 5–7. Shot gain is relative to full $2PGS$ parameter-shift QWC measurement.

Molecule	Method	Full QWC	Groups / param. ↓			Shot gain ↑			No-worse vs Full			Best $\Delta E \downarrow$
		groups/param.	D5	D6	D7	D5	D6	D7	D5	D6	D7	
H2/STO-3G	**Ours**	5	4.65	**4.70**	**4.77**	**7.0%**	**6.0%**	**4.7%**	✓	✗	✗	+0.0008
H2/STO-3G	Random	5	**4.65**	4.70	4.77	**7.0%**	**5.9%**	**4.7%**	✗	✗	✓	+0.0015
HeH+/STO-3G	**Ours**	11	**9.73**	**9.82**	**9.95**	**11.5%**	**10.7%**	**9.6%**	✗	✓	✓	+0.0042
HeH+/STO-3G	Random	11	9.73	9.85	9.95	**11.5%**	**10.5%**	**9.6%**	✗	✗	✗	+0.0207
H3+/STO-3G	**Ours**	19	**16.34**	16.79	**17.05**	**14.0%**	**11.6%**	**10.2%**	✓	✗	✓	+0.0015
H3+/STO-3G	Random	19	16.36	**16.74**	17.06	**13.9%**	**11.9%**	**10.2%**	✗	✗	✗	+0.0481
H4 chain/STO-3G	**Ours**	75	58.99	**60.96**	62.25	**21.3%**	**18.7%**	**17.0%**	✗	✓	✗	−0.0055
H4 chain/STO-3G	Random	75	**58.69**	61.11	**62.13**	**21.8%**	**18.5%**	**17.2%**	✗	✗	✗	+0.0151
LiH/STO-3G	**Ours**	179	72.65	87.48	98.91	**59.4%**	**51.1%**	**44.7%**	✓	✓	✓	−0.0001
LiH/STO-3G	Random	179	**65.92**	**83.27**	**95.42**	**63.2%**	**53.5%**	**46.7%**	✗	✗	✗	+0.6498

E.3 Depth-OOD Baselines by Molecule and Depth

Table 32: **Aggregate depth-OOD baselines at $\tau = 0.997$.** This appendix table summarizes the same experiment as Table 29 across all 15 molecule-depth cells and 60 held-out trajectories. CA seeds count held-out trajectories with $|\Delta E| \leq 1.6$ mHa against Full-QWC. Cells won counts molecule-depth cells where the method has the smallest mean $|\Delta E|$ among all listed methods.

Method	Mean ΔE	Mean $\lvert\Delta E\rvert \downarrow$	CA seeds	Shot gain ↑	Groups/param.	Cells won
Ours	**-0.00118**	**0.00716**	**18/60**	16.2%	38.90	**12/15**
Global coeff. same budget	+0.00226	0.01139	11/60	16.2%	38.88	3/15
Random same budget	+0.25163	0.26431	2/60	16.7%	38.13	0/15
TITAN (Peng et al.)	+0.05683	0.09985	3/60	41.1%	24.45	0/15
TITAN+**Ours**	+0.05647	0.09981	4/60	**48.6%**	16.52	0/15

Table 33: **Molecule-depth detail for depth-OOD baselines at $\tau = 0.997$.** Each entry reports mean $|\Delta E|$ in Hartree followed by shot saving. Rows are grouped by molecular Hamiltonian and unseen evaluation depth. The model is trained on depths 1–4 and evaluated on depths 5–7 with four held-out seeds per molecule-depth cell. Green bold marks the lowest mean $|\Delta E|$ in each row.

Molecule	Depth	**Ours**	Global coeff.	TITAN (Peng et al.)	TITAN+**Ours**	Random
H2 (4q)	5	**0.00119 / 5.7%**	0.00786 / 5.7%	0.03143 / 32.0%	0.03200 / 34.3%	0.01792 / 5.7%
H2 (4q)	6	**0.00101 / 4.8%**	0.00915 / 4.7%	0.01577 / 32.8%	0.01678 / 34.6%	0.02628 / 4.7%
H2 (4q)	7	**0.00199 / 3.8%**	0.00352 / 3.8%	0.01781 / 29.5%	0.01761 / 31.3%	0.00650 / 3.8%
HeH+ (4q)	5	**0.00572 / 8.4%**	0.01661 / 8.4%	0.08943 / 32.7%	0.08406 / 37.9%	0.06841 / 8.5%
HeH+ (4q)	6	**0.00484 / 7.6%**	0.01717 / 7.6%	0.11144 / 30.5%	0.11416 / 34.4%	0.03489 / 7.5%
HeH+ (4q)	7	0.01149 / 6.7%	**0.00857 / 6.7%**	0.09757 / 31.3%	0.09613 / 34.6%	0.07671 / 6.6%
H3+ (6q)	5	**0.01204 / 10.6%**	0.01437 / 10.7%	0.05297 / 38.7%	0.05186 / 44.9%	0.05029 / 10.7%
H3+ (6q)	6	**0.00422 / 8.7%**	0.01294 / 8.6%	0.04326 / 28.1%	0.04484 / 33.8%	0.02628 / 8.7%
H3+ (6q)	7	**0.00378 / 7.4%**	0.00473 / 7.4%	0.04070 / 31.6%	0.04348 / 35.8%	0.01635 / 7.3%
H4 chain (8q)	5	**0.02352 / 17.1%**	0.04048 / 17.0%	0.12680 / 42.3%	0.12740 / 52.2%	0.15799 / 17.2%
H4 chain (8q)	6	**0.00929 / 14.6%**	0.01119 / 14.6%	0.05705 / 39.4%	0.05890 / 47.6%	0.07854 / 14.6%
H4 chain (8q)	7	**0.00614 / 13.0%**	0.00704 / 13.1%	0.04519 / 36.1%	0.04317 / 43.9%	0.05636 / 13.0%
LiH (12q)	5	0.00877 / 52.9%	**0.00208 / 52.9%**	0.22542 / 73.4%	0.22498 / 93.2%	2.18927 / 55.4%
LiH (12q)	6	**0.01041 / 44.3%**	0.01414 / 44.4%	0.34075 / 71.2%	0.33967 / 88.5%	0.74399 / 46.3%
LiH (12q)	7	0.00307 / 37.9%	**0.00095 / 38.0%**	0.20216 / 66.6%	0.20205 / 82.5%	0.41488 / 39.7%

F CONTROLLED STRUCTURAL OOD TRANSFER SUMMARY

Figure 3 presents the matched-budget energy comparison compactly; Table 34 gives the corresponding exact energy gaps, success counts, shot gains, and selected-group budgets.

Table 34: **Controlled OOD transfer at $\tau = 0.997$. (a)** Each target Hamiltonian is absent from training; no-worse counts final energies at most 1.6 mHa above Full-QWC, accepting lower energies. **(b)** Each named ansatz is unseen during training; strict CA counts seeds with $|\Delta E| \leq 1.6$ mHa, and our method wins counts molecule–depth cells with a smaller mean absolute gap than Random. In both panels, Random uses the same parameter-specific QWC-group count as our method, isolating relevance ordering at a matched measurement budget. Within each paired setting, bold marks the lower mean $|\Delta E|$ and larger success count; underline marks the runner-up. Budget columns are intentionally unranked. Complete TITAN controls are Tables 38 and 41 in Appendices G and H.

(a) Hamiltonian-family OOD

Test Hamiltonian	Method	Full groups	Mean ΔE	Mean $\lvert\Delta E\rvert \downarrow$	No-worse vs Full	Shot gain ↑	Groups/param.
TFIM $h = 0.7$	**Ours**	2	-1.55×10^{-15}	3.66×10^{-15}	**16/16**	6.13%	1.877
TFIM $h = 0.7$	Random	2	-0.00253	<u>0.00710</u>	<u>6/16</u>	6.20%	1.876
TFIM $h = 1.2$	**Ours**	2	0.21181	<u>0.22777</u>	<u>0/16</u>	13.02%	1.740
TFIM $h = 1.2$	Random	2	0.03008	0.03988	**1/16**	13.02%	1.740
Heisenberg XXZ+field	**Ours**	3	-0.00011	0.00011	**16/16**	8.69%	2.739
Heisenberg XXZ+field	Random	3	0.37128	<u>0.43000</u>	<u>0/16</u>	8.69%	2.739
Heisenberg J1–J2	**Ours**	3	0.00563	0.00567	**14/16**	8.72%	2.738
Heisenberg J1–J2	Random	3	0.22592	<u>0.30762</u>	<u>0/16</u>	8.70%	2.739
All targets	**Ours**	–	0.05433	0.05839	**46/64**	9.14%	2.274
All targets	Random	–	0.15624	<u>0.19615</u>	<u>7/64</u>	9.15%	2.273

(b) Named-ansatz OOD

OOD ansatz	Method	Mean ΔE	Mean $\lvert\Delta E\rvert \downarrow$	Strict CA seeds	Shot gain ↑	**Ours** wins
EfficientSU2	**Ours**	$+0.0030$	**0.0119**	**13/60**	9.5%	–
EfficientSU2	Random	$+0.0344$	<u>0.0630</u>	<u>2/60</u>	9.4%	14/15
HEA-RYRZ-Circular	**Ours**	$+0.0033$	**0.0088**	**12/60**	12.9%	–
HEA-RYRZ-Circular	Random	$+0.0627$	<u>0.0690</u>	<u>2/60</u>	13.2%	15/15
RealAmplitudes	**Ours**	-0.0006	**0.0030**	**29/60**	13.5%	–
RealAmplitudes	Random	$+0.1867$	<u>0.2182</u>	<u>1/60</u>	13.8%	15/15
StronglyEntanglingLayers	**Ours**	$+0.0395$	<u>0.0453</u>	<u>0/60</u>	11.1%	–
StronglyEntanglingLayers	Random	$+0.0271$	**0.0390**	**2/60**	11.2%	4/15

G MOLECULE- AND HAMILTONIAN-FAMILY OOD TRANSFER

The compact comparison between our method and Random on Hamiltonian transfer appears in panel (a) of Table 34 in Appendix F; Table 38 below preserves the corresponding TITAN controls.

G.1 CLEAN MOLECULE-OOD HAMILTONIAN TRANSFER

For each split, one molecule is held out and the relevance model is trained on the remaining four molecules at depths 1–4. The matched TITAN comparison uses the corresponding holdout-specific checkpoint, so neither our method nor either dagger composition has seen the testing molecule.

Table 35: **Leave-one-molecule-out Hamiltonian transfer at $\tau = 0.997$.** Representative depths 1 and 4 are shown; depths 2 and 3 appear in Appendix Table 37. TITAN† uses a cell-wise freezing threshold selected to match our method's shot gain; TITAN†+**Ours** uses the same active-parameter mask and the matching holdout-specific selector. Green/red ΔE and ✓/✗ follow the one-sided 1.6 mHa no-worse criterion. Bold marks the closest mean ΔE to Full; bold-underlined shot gain marks the largest saving.

Molecule	Depth	Metric	**Ours**	Random	TITAN (Peng et al.)	TITAN+**Ours**	TITAN†	TITAN†+**Ours**
H2 (4q)	Depth 1	ΔE vs Full ↓	**+0.0016**	+0.3323	+0.0228	+0.0258	+0.0262	+0.0272
	Depth 1	Shot gain ↑	55.2%	56.1%	61.8%	**88.8%**	56.0%	86.0%
	Depth 1	Chem. acc.	✗	✗	✗	✗	✗	✗
	Depth 4	ΔE vs Full ↓	**-0.0002**	+0.0249	**-0.0003**	**-0.0029**	**+0.0010**	**+0.0007**
	Depth 4	Shot gain ↑	9.9%	9.7%	40.8%	**45.3%**	10.2%	16.0%
	Depth 4	Chem. acc.	✓	✗	✓	✓	✓	✓
HeH+ (4q)	Depth 1	ΔE vs Full ↓	**+0.0021**	+0.5282	+0.1040	+0.1063	+0.0090	+0.0114
	Depth 1	Shot gain ↑	32.2%	30.4%	56.2%	**75.6%**	30.9%	58.3%
	Depth 1	Chem. acc.	✗	✗	✗	✗	✗	✗
	Depth 4	ΔE vs Full ↓	**-0.0022**	+0.0452	**-0.0642**	**-0.0650**	+0.0111	+0.0112
	Depth 4	Shot gain ↑	8.1%	8.1%	38.3%	**42.7%**	4.7%	12.5%
	Depth 4	Chem. acc.	✓	✗	✓	✓	✗	✗
H3+ (6q)	Depth 1	ΔE vs Full ↓	**+0.0051**	+0.9872	+0.0871	+0.0867	+0.0871	+0.0867
	Depth 1	Shot gain ↑	56.2%	60.6%	58.2%	**94.3%**	58.2%	94.3%
	Depth 1	Chem. acc.	✗	✗	✗	✗	✗	✗
	Depth 4	ΔE vs Full ↓	**-0.0052**	+0.0504	+0.0116	+0.0128	+0.0049	**+0.0017**
	Depth 4	Shot gain ↑	10.0%	10.0%	45.9%	**52.0%**	6.2%	16.0%
	Depth 4	Chem. acc.	✓	✗	✗	✗	✗	✗
H4 chain (8q)	Depth 1	ΔE vs Full ↓	**+0.0038**	+1.5543	+0.0220	+0.0236	+0.0220	+0.0236
	Depth 1	Shot gain ↑	63.7%	68.3%	62.7%	**97.4%**	62.7%	97.4%
	Depth 1	Chem. acc.	✗	✗	✗	✗	✗	✗
	Depth 4	ΔE vs Full ↓	**+0.0652**	+0.3159	+0.1904	+0.1887	+0.0770	+0.0752
	Depth 4	Shot gain ↑	28.5%	29.4%	44.7%	**63.2%**	28.6%	50.4%
	Depth 4	Chem. acc.	✗	✗	✗	✗	✗	✗
LiH (12q)	Depth 1	ΔE vs Full ↓	**+0.0022**	+0.4404	**-0.0881**	**-0.0880**	+0.0387	+0.0385
	Depth 1	Shot gain ↑	40.4%	37.9%	71.9%	**88.5%**	47.1%	71.9%
	Depth 1	Chem. acc.	✗	✗	✓	✓	✗	✗
	Depth 4	ΔE vs Full ↓	**-0.0023**	+0.2301	**-0.0100**	**-0.0092**	+0.0516	+0.0510
	Depth 4	Shot gain ↑	17.9%	17.7%	76.2%	**81.9%**	20.2%	33.8%
	Depth 4	Chem. acc.	✓	✗	✓	✓	✗	✗

Table 35 directly compares group selection and parameter freezing under Hamiltonian shift. The dagger columns control the TITAN-only shot budget cell by cell, while the final column measures the additional savings from composing the matched active-parameter mask with learned QWC-group selection.

G.2 Leave-One-Molecule-Out Uncertainty and Intermediate Depths

Table 36: **Uncertainty view for the leave-one-molecule-out comparison at $\tau = 0.997$.** This appendix table reports mean $\Delta E \pm$ s.d. for the representative depths 1 and 4 shown in Table 35. Green/red coloring follows the mean ΔE rule used in Table 35.

Molecule	Depth	**Ours**	Random	TITAN (Peng et al.)	TITAN+**Ours**
H2 (4q)	Depth 1	$+0.0016 \pm 0.0035$	$+0.3323 \pm 0.0518$	$+0.0228 \pm 0.0259$	$+0.0258 \pm 0.0255$
	Depth 4	-0.0002 ± 0.0099	$+0.0249 \pm 0.0284$	-0.0003 ± 0.0206	-0.0029 ± 0.0213
HeH+ (4q)	Depth 1	$+0.0021 \pm 0.0034$	$+0.5282 \pm 0.1378$	$+0.1040 \pm 0.0586$	$+0.1063 \pm 0.0578$
	Depth 4	-0.0022 ± 0.0065	$+0.0452 \pm 0.0291$	-0.0642 ± 0.0614	-0.0650 ± 0.0628
H3+ (6q)	Depth 1	$+0.0051 \pm 0.0041$	$+0.9872 \pm 0.0979$	$+0.0871 \pm 0.0466$	$+0.0867 \pm 0.0453$
	Depth 4	-0.0052 ± 0.0072	$+0.0504 \pm 0.0457$	$+0.0116 \pm 0.0819$	$+0.0128 \pm 0.0771$
H4 chain (8q)	Depth 1	$+0.0038 \pm 0.0051$	$+1.5543 \pm 0.1952$	$+0.0220 \pm 0.0193$	$+0.0236 \pm 0.0171$
	Depth 4	$+0.0652 \pm 0.0666$	$+0.3159 \pm 0.1015$	$+0.1904 \pm 0.0841$	$+0.1887 \pm 0.0819$
LiH (12q)	Depth 1	$+0.0022 \pm 0.0031$	$+0.4404 \pm 0.1245$	-0.0881 ± 0.0470	-0.0880 ± 0.0466
	Depth 4	-0.0023 ± 0.0049	$+0.2301 \pm 0.1702$	-0.0100 ± 0.2342	-0.0092 ± 0.2339

Table 37: **Appendix depth-2/3 view for leave-one-molecule-out Hamiltonian transfer at $\tau = 0.997$.** This table uses the same protocol and layout as Table 35: one molecule is held out for testing, the remaining four molecules train the selector, and depths are kept fixed between train and test. ΔE entries report mean $\pm$ s.d. across held-out seeds. ✓ denotes the one-sided no-worse chemical-accuracy criterion $\Delta E \leq 1.6$ mHa relative to Full-QWC training; ✗ denotes a larger positive energy gap. In ΔE rows, green means $\Delta E \leq 1.6$ mHa (including negative gaps) and red means a larger positive gap.

Molecule	Depth	Metric	**Ours**	Random	TITAN (Peng et al.)	TITAN+**Ours**
H2 (4q)	Depth 2	ΔE vs Full ↓	$+0.0046 \pm 0.0063$	$+0.1387 \pm 0.1420$	-0.0200 ± 0.1784	-0.0180 ± 0.1775
	Depth 2	Shot gain ↑	29.9%	30.1%	54.0%	**72.3%**
	Depth 2	Chem. acc.	✗	✗	✓	✓
	Depth 3	ΔE vs Full ↓	$+0.0576 \pm 0.1057$	$+0.0814 \pm 0.0453$	$+0.0223 \pm 0.0453$	$+0.0205 \pm 0.0409$
	Depth 3	Shot gain ↑	19.3%	21.3%	39.7%	**52.0%**
	Depth 3	Chem. acc.	✗	✗	✗	✗
HeH+ (4q)	Depth 2	ΔE vs Full ↓	$+0.0167 \pm 0.0183$	$+0.2844 \pm 0.0547$	-0.1194 ± 0.1021	-0.1191 ± 0.1008
	Depth 2	Shot gain ↑	22.2%	23.7%	54.4%	**70.5%**
	Depth 2	Chem. acc.	✗	✗	✓	✓
	Depth 3	ΔE vs Full ↓	$+0.0133 \pm 0.0145$	$+0.0677 \pm 0.1411$	$+0.0112 \pm 0.0424$	$+0.0102 \pm 0.0444$
	Depth 3	Shot gain ↑	13.2%	13.6%	43.3%	**52.2%**
	Depth 3	Chem. acc.	✗	✗	✗	✗
H3+ (6q)	Depth 2	ΔE vs Full ↓	$+0.0013 \pm 0.0018$	$+0.3811 \pm 0.0782$	-0.0031 ± 0.0610	-0.0036 ± 0.0629
	Depth 2	Shot gain ↑	27.6%	27.3%	50.0%	**71.0%**
	Depth 2	Chem. acc.	✓	✗	✓	✓
	Depth 3	ΔE vs Full ↓	$+0.0017 \pm 0.0076$	$+0.1683 \pm 0.1216$	$+0.0338 \pm 0.0777$	$+0.0332 \pm 0.0800$
	Depth 3	Shot gain ↑	16.2%	16.1%	47.1%	**57.7%**
	Depth 3	Chem. acc.	✗	✗	✗	✗
H4 chain (8q)	Depth 2	ΔE vs Full ↓	$+0.0065 \pm 0.0112$	$+1.1356 \pm 0.1572$	$+0.1521 \pm 0.1283$	$+0.1556 \pm 0.1269$
	Depth 2	Shot gain ↑	52.3%	57.1%	59.1%	**88.9%**
	Depth 2	Chem. acc.	✗	✗	✗	✗
	Depth 3	ΔE vs Full ↓	$+0.0107 \pm 0.0102$	$+0.3099 \pm 0.0388$	-0.0650 ± 0.1313	-0.0615 ± 0.1294
	Depth 3	Shot gain ↑	39.3%	40.6%	48.3%	**72.9%**
	Depth 3	Chem. acc.	✗	✗	✓	✓
LiH (12q)	Depth 2	ΔE vs Full ↓	$+0.0026 \pm 0.0026$	$+1.0068 \pm 0.1733$	$+0.0806 \pm 0.0923$	$+0.0803 \pm 0.0926$
	Depth 2	Shot gain ↑	32.1%	30.6%	77.1%	**87.1%**
	Depth 2	Chem. acc.	✗	✗	✗	✗
	Depth 3	ΔE vs Full ↓	$+0.0038 \pm 0.0133$	$+0.5742 \pm 0.2506$	$+0.3972 \pm 0.2246$	$+0.3975 \pm 0.2246$
	Depth 3	Shot gain ↑	24.1%	23.6%	76.4%	**84.2%**
	Depth 3	Chem. acc.	✗	✗	✗	✗

G.3 Full Hamiltonian-Family Baseline Comparison

This Hamiltonian-transfer experiment asks whether the learned relevance selector transfers beyond molecular electronic-structure Hamiltonians. We train one shared selector on five PySCF/OpenFermion molecular Hamiltonians (H2, HeH^+, $H3^+$, H4 chain, and LiH; STO-3G basis) plus one 8-qubit Heisenberg-chain Hamiltonian, using ansatz depths 1–4. We then evaluate closed-loop online VQE at the same depths on four unseen spin Hamiltonians: two transverse-field Ising models (TFIM) and two Heisenberg variants. The ansatz is the same layered RY/RZ circuit with a linear entangler, and the top-mass threshold is fixed at $\tau = 0.997$.

Table 38: **Full Hamiltonian-family OOD baseline comparison at $\tau = 0.997$.** The selector is trained on molecular Hamiltonians plus one Heisenberg Hamiltonian and evaluated on unseen TFIM/Heisenberg-family Hamiltonians at the same depths 1–4. This appendix table preserves the TITAN and TITAN+**Ours** controls omitted from the compact comparison in Table 34. No-worse vs Full reports seeds whose final energy is not more than 1.6 mHa above full-QWC training, accepting lower energies. Full groups is the number of QWC groups in the target Hamiltonian. Bold green marks the lowest mean $|\Delta E|$ within each test Hamiltonian; underlined green marks the largest shot gain.

Test Hamiltonian	Method	Full groups	Mean ΔE	Mean $\lvert\Delta E\rvert \downarrow$	No-worse vs Full	Shot gain ↑	Groups/param.
TFIM $h = 0.7$	**Ours**	2	-1.55×10^{-15}	3.66×10^{-15}	**16/16**	6.13%	1.877
TFIM $h = 0.7$	Random	2	−0.00253	0.00710	6/16	6.20%	1.876
TFIM $h = 0.7$	TITAN (Peng et al., 2026)	2	−0.00587	0.01964	1/16	56.11%	0.878
TFIM $h = 0.7$	TITAN+**Ours**	2	−0.00587	0.01964	1/16	**57.37%**	0.853
TFIM $h = 1.2$	**Ours**	2	0.21181	0.22777	0/16	13.02%	1.740
TFIM $h = 1.2$	Random	2	0.03008	0.03988	1/16	13.02%	1.740
TFIM $h = 1.2$	TITAN (Peng et al., 2026)	2	0.02293	0.05889	1/16	49.22%	1.016
TFIM $h = 1.2$	TITAN+**Ours**	2	0.22750	0.24289	0/16	**53.81%**	0.924
Heisenberg XXZ+field	**Ours**	3	−0.00011	0.00011	**16/16**	8.69%	2.739
Heisenberg XXZ+field	Random	3	0.37128	0.43000	0/16	8.69%	2.739
Heisenberg XXZ+field	TITAN (Peng et al., 2026)	3	0.50014	0.71973	0/16	45.31%	1.641
Heisenberg XXZ+field	TITAN+**Ours**	3	0.49967	0.71926	0/16	**46.79%**	1.596
Heisenberg J1–J2	**Ours**	3	0.00563	0.00567	**14/16**	8.72%	2.738
Heisenberg J1–J2	Random	3	0.22592	0.30762	0/16	8.70%	2.739
Heisenberg J1–J2	TITAN (Peng et al., 2026)	3	0.58258	0.67115	1/16	44.13%	1.676
Heisenberg J1–J2	TITAN+**Ours**	3	0.58165	0.67070	1/16	**45.64%**	1.631
All targets	**Ours**	–	0.05433	0.05839	**46/64**	9.14%	2.274
All targets	Random	–	0.15624	0.19615	7/64	9.15%	2.273
All targets	TITAN (Peng et al., 2026)	–	0.27494	0.36735	3/64	48.69%	1.303
All targets	TITAN+**Ours**	–	0.32574	0.41312	2/64	**50.90%**	1.251

The result separates Hamiltonian-family transfer from ordinary molecular holdout. Our method remains the most accurate method on three of the four unseen spin Hamiltonians, including both Heisenberg variants and TFIM at $h = 0.7$. In contrast, TITAN-style parameter freezing obtains much larger shot savings but usually changes the optimization trajectory enough to lose final-energy fidelity. TFIM at $h = 1.2$ is the failure case: the learned molecular-plus-Heisenberg relevance ordering is not yet robust to this field regime. This suggests that Hamiltonian-family coverage, not just more molecular seeds, is the next dataset axis for a stronger general-purpose selector.

G.4 HAMILTONIAN-PANEL STRESS TEST

This expanded panel systematically tests Hamiltonian-family OOD transfer. We train one shared selector on five molecular Hamiltonians plus two 8-qubit Heisenberg-chain Hamiltonians, using depths 1–4 and the same layered RY/RZ ansatz. We then evaluate the saved model on held-out transverse-field Ising models, unseen Heisenberg variants, and a larger-group JW-like parity spin Hamiltonian at depths 1–4 with $\tau = 0.997$. The baseline comparison adds cell-wise cost-matched TITAN controls so that parameter freezing and QWC-group selection are compared at similar measurement budgets.

Table 39: **Hamiltonian-panel transfer at $\tau = 0.997$.** Each testing Hamiltonian is held out from training; representative depths 1 and 4 are shown. TITAN† uses the cell-wise freezing threshold closest to our method's shot gain, and TITAN†**+Ours** reuses that active-parameter mask before group selection. Green/red ΔE and ✓/✗ follow the one-sided 1.6 mHa no-worse criterion. Bold marks the closest mean ΔE to Full; bold-underlined shot gain marks the largest saving.

Testing Hamiltonian	Depth	Metric	**Ours**	Random	TITAN (Peng et al.)	TITAN+**Ours**	TITAN†	TITAN†**+Ours**
TFIM $h = 0.5$	Depth 1	ΔE vs Full ↓	**+0.0000**	**-0.0043**	**-0.0139**	**-0.0139**	**+0.0000**	**+0.0000**
	Depth 1	Shot gain ↑	2.0%	2.0%	56.8%	**57.5%**	0.0%	2.0%
	Depth 1	Chem. acc.	✓	✓	✓	✓	✓	✓
	Depth 4	ΔE vs Full ↓	**-0.0002**	**-0.0008**	+0.0062	+0.0063	**+0.0000**	**-0.0002**
	Depth 4	Shot gain ↑	2.6%	2.6%	59.0%	**59.3%**	0.0%	2.6%
	Depth 4	Chem. acc.	✓	✓	✗	✗	✓	✓
TFIM $h = 0.7$	Depth 1	ΔE vs Full ↓	**+0.0000**	**-0.0138**	**-0.0227**	**-0.0227**	**+0.0000**	**+0.0000**
	Depth 1	Shot gain ↑	2.9%	2.9%	56.4%	**57.6%**	0.0%	2.9%
	Depth 1	Chem. acc.	✓	✓	✓	✓	✓	✓
	Depth 4	ΔE vs Full ↓	**+0.0000**	**-0.0018**	+0.0107	+0.0107	**+0.0000**	**+0.0000**
	Depth 4	Shot gain ↑	2.4%	2.4%	53.0%	**53.5%**	0.0%	2.4%
	Depth 4	Chem. acc.	✓	✓	✗	✗	✓	✓
TFIM $h = 0.9$	Depth 1	ΔE vs Full ↓	+0.0771	**-0.0010**	**-0.0018**	+0.0784	**-0.0002**	+0.0771
	Depth 1	Shot gain ↑	25.0%	25.0%	54.4%	**59.8%**	27.1%	37.6%
	Depth 1	Chem. acc.	✗	✓	✓	✗	✓	✗
	Depth 4	ΔE vs Full ↓	+0.0302	+0.0064	+0.0028	+0.0351	**-0.0012**	+0.0306
	Depth 4	Shot gain ↑	6.2%	6.2%	47.1%	**49.0%**	6.0%	11.1%
	Depth 4	Chem. acc.	✗	✗	✗	✗	✓	✗
Heis. XXZ+X field	Depth 1	ΔE vs Full ↓	**+0.0000**	+0.8893	+1.2891	+1.2891	+0.2391	+0.2391
	Depth 1	Shot gain ↑	16.7%	16.7%	52.2%	**53.5%**	19.4%	29.6%
	Depth 1	Chem. acc.	✓	✗	✗	✗	✗	✗
	Depth 4	ΔE vs Full ↓	**+0.0000**	+0.3996	+0.5969	+0.5969	**-0.1146**	**-0.1146**
	Depth 4	Shot gain ↑	4.2%	4.2%	36.3%	**38.1%**	3.7%	7.7%
	Depth 4	Chem. acc.	✓	✗	✗	✗	✓	✓
Heis. J1–J2	Depth 1	ΔE vs Full ↓	**+0.0000**	+0.6602	+0.9835	+0.9835	+0.0863	+0.0863
	Depth 1	Shot gain ↑	16.7%	16.7%	50.6%	**52.2%**	12.7%	25.1%
	Depth 1	Chem. acc.	✓	✗	✗	✗	✗	✗
	Depth 4	ΔE vs Full ↓	**+0.0000**	**-0.0010**	+0.5678	+0.5678	+0.1550	+0.1550
	Depth 4	Shot gain ↑	4.2%	4.2%	34.2%	**36.0%**	3.5%	7.4%
	Depth 4	Chem. acc.	✓	✓	✗	✗	✗	✗
Heis. XXY+Z field	Depth 1	ΔE vs Full ↓	**+0.0000**	+0.9213	+0.8904	+0.8904	+0.1382	+0.1382
	Depth 1	Shot gain ↑	16.7%	16.7%	49.6%	**51.9%**	18.9%	29.4%
	Depth 1	Chem. acc.	✓	✗	✗	✗	✗	✗
	Depth 4	ΔE vs Full ↓	**+0.0000**	+0.0742	**-0.0137**	**-0.0137**	+0.0219	+0.0219
	Depth 4	Shot gain ↑	4.2%	4.2%	31.1%	**33.3%**	3.6%	7.5%
	Depth 4	Chem. acc.	✓	✗	✓	✓	✗	✗
JW-like parity spin	Depth 1	ΔE vs Full ↓	+0.0312	+0.6318	+0.1336	+0.1387	**+0.0049**	+0.0350
	Depth 1	Shot gain ↑	35.7%	35.6%	55.5%	**76.9%**	31.4%	59.5%
	Depth 1	Chem. acc.	✗	✗	✗	✗	✗	✗
	Depth 4	ΔE vs Full ↓	**-0.0046**	+0.0076	+0.0822	+0.0738	**+0.0012**	**-0.0198**
	Depth 4	Shot gain ↑	11.6%	11.5%	45.7%	**52.1%**	9.8%	19.6%
	Depth 4	Chem. acc.	✓	✗	✗	✗	✓	✓

Table 39 shows a useful but bounded generalization result. Our method is full-QWC-equivalent on the held-out Heisenberg variants and low-field TFIM, whereas shot-matched parameter freezing still incurs large positive drift on most Heisenberg cells. On low-field TFIM, our method saves only 2–3%, so the nearest TITAN-only operating point is no freezing; the composed dagger column then reduces exactly to our method group selection. The JW-like target remains a harder family shift, although matched TITAN substantially narrows the gap relative to aggressive freezing. These controls show that the proposed method's stability is not explained by a more conservative parameter-freezing threshold.

G.5 JOINT MOLECULE+DEPTH OOD STRESS TEST

This joint-OOD experiment is preserved as a strict stress test rather than the clean molecule-OOD experiment. The model is trained on H2, HeH+, H3+, and LiH at RY/RZ depths 1–4, then evaluated without retraining on H4 chain at unseen depths 5–7. Thus both the molecular Hamiltonian and the ansatz depth are held out. We do not claim this is the first cross-molecule VQE transfer experiment broadly: prior work such as Meta-VQE (Cervera-Lierta et al., 2021) studies learning over parameterized Hamiltonians. Here the role of this appendix result is narrower: it probes whether a learned parameter-by-QWC measurement policy survives simultaneous Hamiltonian and depth shift.

Table 40: **Joint molecule+depth OOD transfer to H4 at $\tau = 0.997$.** The relevance model is trained on H2, HeH+, H3+, and LiH with RY/RZ depths 1–4 and evaluated on held-out H4 chain at unseen depths 5–7 with seeds 97/101/107/109. Each depth cell reports mean ΔE versus full-QWC followed by mean shot saving. Mean $|\Delta E|$ averages the three H4 depths. No-worse counts held-out seeds with $\Delta E \leq 1.6$ mHa, so lower final energy is accepted. Global coefficient uses the same per-parameter group count as **Ours** but ranks QWC groups by Hamiltonian coefficient magnitude.

Method	Depth 5	Depth 6	Depth 7	Mean $\|\Delta E\|$ ↓	No-worse seeds	Shot gain ↑
Ours	+0.0342 / **23.5%**	+0.0030 / 20.3%	+0.0060 / 18.8%	0.0251	**6/12**	20.9%
Random	+0.1682 / 23.6%	+0.1171 / 20.4%	+0.0322 / 18.9%	0.1064	1/12	21.0%
Global coeff.	+0.0380 / 23.8%	−0.0020 / **20.2%**	+0.0049 / **18.9%**	**0.0201**	5/12	20.9%
TITAN (Peng et al.)	+0.1011 / 42.3%	+0.0070 / 39.4%	−0.0289 / 36.1%	0.0763	5/12	39.3%
TITAN+**Ours**	+0.1040 / 56.8%	+0.0071 / 52.7%	−0.0300 / 48.3%	0.0791	5/12	**52.6%**

Table 40 shows that joint molecule+depth OOD transfer is substantially harder than depth-OOD within the same molecule panel. Our method is much more stable than matched-budget random selection and wins the hardest H4 depth-5 cell, but the static Global coefficient baseline is competitive on depths 6–7 and obtains the lowest aggregate mean absolute gap. We therefore treat this joint shift as a stress test and diagnostic rather than the clean molecule-OOD result: the current learner captures useful transferable structure, but stronger Hamiltonian/QWC group features are needed for robust cross-Hamiltonian generalization.

H ANSATZ-FAMILY OOD TRANSFER

The compact comparison between our method and Random on named-ansatz transfer appears in panel (b) of Table 34 in Appendix F; Table 41 below preserves the corresponding cost-matched TITAN controls.

H.1 FULL NAMED-ANSATZ ZERO-SHOT BASELINES

The preceding depth-OOD experiments keep the ansatz family fixed and increase depth. We next test a harder transfer axis: the saved depth-OOD model is trained only on the HEA-RYRZ-linear ansatz at depths 1–4, then evaluated *zero-shot* on named ansatz families at depths 5–7. The molecule panel, held-out seeds, top-mass threshold, and online VQE protocol are unchanged: H2, HeH+, H3+, H4 chain, and LiH in STO-3G; four held-out seeds per molecule-depth cell; $\tau = 0.997$; and 1024 shots per parameter-shift measurement setting.

Table 41: **Full zero-shot baseline comparison on named ansatz families.** The saved depth-OOD selector is trained on HEA-RYRZ-linear depths 1–4 and evaluated without retraining on depths 5–7, five STO-3G molecules, and four held-out seeds per molecule-depth cell at $\tau = 0.997$. This appendix table preserves the TITAN† and TITAN†**+Ours** controls omitted from the compact comparison in Table 34. Random uses the same per-parameter group budget as our method. TITAN† directly chooses the active-parameter count in each molecule-depth cell so its shot gain matches our method; TITAN†**+Ours** applies group sparsification inside that active set. Strict CA counts seeds with $|\Delta E| \leq 1.6$ mHa. The final column reports cells, out of 15, where our method has a smaller mean absolute gap than the method in that row.

OOD ansatz	Method	Mean ΔE	Mean $\lvert\Delta E\rvert \downarrow$	Strict CA seeds	Shot gain ↑	**Ours** wins
EfficientSU2	**Ours**	+0.0030	0.0119	13/60	9.5%	–
EfficientSU2	Random	+0.0344	0.0630	2/60	9.4%	14/15
EfficientSU2	TITAN†	−0.0048	**0.0118**	**14/60**	9.5%	7/15
EfficientSU2	TITAN†**+Ours**	−0.0010	0.0175	6/60	**17.4%**	9/15
HEA-RYRZ-Circular	**Ours**	+0.0033	**0.0088**	12/60	12.9%	–
HEA-RYRZ-Circular	Random	+0.0627	0.0690	2/60	13.2%	15/15
HEA-RYRZ-Circular	TITAN†	−0.0012	0.0197	**14/60**	13.0%	9/15
HEA-RYRZ-Circular	TITAN†**+Ours**	+0.0011	0.0209	8/60	**23.0%**	11/15
RealAmplitudes	**Ours**	−0.0006	**0.0030**	**29/60**	13.5%	–
RealAmplitudes	Random	+0.1867	0.2182	1/60	13.8%	15/15
RealAmplitudes	TITAN†	+0.0151	0.0429	8/60	13.5%	15/15
RealAmplitudes	TITAN†**+Ours**	+0.0155	0.0435	5/60	**23.4%**	15/15
StronglyEntanglingLayers	**Ours**	+0.0395	0.0453	0/60	11.1%	–
StronglyEntanglingLayers	Random	+0.0271	0.0390	2/60	11.2%	4/15
StronglyEntanglingLayers	TITAN†	−0.0032	**0.0080**	**34/60**	11.1%	0/15
StronglyEntanglingLayers	TITAN†**+Ours**	+0.0343	0.0407	0/60	**20.0%**	5/15

Table 41 shows that the learned relevance ordering transfers well to EfficientSU2, HEA-RYRZ-Circular, and RealAmplitudes, where our method beats Random on 44 of 45 cells. Cost matching rules out a universal advantage from parameter freezing: TITAN† is comparable on EfficientSU2, but our method has smaller gaps on 9/15 HEA-RYRZ-Circular cells and all 15 RealAmplitudes cells. The ordering reverses on StronglyEntanglingLayers, where TITAN† reaches a 0.0080 Ha mean absolute gap while our method fails. Composing the controls adds 8.9–10.0 percentage points of shot saving beyond the matched TITAN arm, but generally increases energy error. Thus parameter freezing and group selection are complementary cost axes, while their reliability remains ansatz dependent.

H.2 Predictor Architecture Under Named-Ansatz OOD

Table 42: **Predictor comparison under named-ansatz OOD at $\tau = 0.997$.** Transformer denotes the depth-OOD predictor used by our method in Table 34. MLP, ResNet-18, and GNN use the same labels, token features, train/evaluation splits, and online VQE protocol; only the relevance predictor changes. The final column counts cells, out of 15, where Transformer has a smaller mean absolute gap than the method in that row. Boldface marks the smallest mean absolute gap and largest strict-CA count per ansatz; bold-underlined values mark the largest shot gain.

OOD ansatz	Predictor	Mean ΔE	Mean $\lvert\Delta E\rvert \downarrow$	Strict CA seeds	Shot gain ↑	Transformer wins
EfficientSU2	Transformer	+0.0030	0.0119	13/60	9.5%	–
EfficientSU2	MLP	+0.0005	0.0088	9/60	8.7%	6/15
EfficientSU2	ResNet-18	−0.0003	**0.0066**	**23/60**	**11.0%**	3/15
EfficientSU2	GNN	+0.0009	0.0089	22/60	5.1%	2/15
HEA-RYRZ-Circular	Transformer	+0.0033	0.0088	12/60	**12.9%**	–
HEA-RYRZ-Circular	MLP	+0.0038	0.0102	15/60	12.0%	8/15
HEA-RYRZ-Circular	ResNet-18	+0.0014	**0.0048**	**31/60**	12.7%	4/15
HEA-RYRZ-Circular	GNN	+0.0038	0.0094	16/60	10.7%	5/15
RealAmplitudes	Transformer	−0.0006	**0.0030**	**29/60**	13.5%	–
RealAmplitudes	MLP	−0.0033	0.0106	17/60	**17.6%**	14/15
RealAmplitudes	ResNet-18	−0.0020	0.0080	23/60	15.5%	9/15
RealAmplitudes	GNN	+0.0031	0.0149	9/60	17.0%	14/15
StronglyEntanglingLayers	Transformer	+0.0395	0.0453	0/60	11.1%	–
StronglyEntanglingLayers	MLP	+0.0148	0.0243	5/60	11.0%	1/15
StronglyEntanglingLayers	ResNet-18	−0.0002	**0.0073**	**27/60**	**11.3%**	0/15
StronglyEntanglingLayers	GNN	+0.0018	0.0116	10/60	9.8%	0/15

ResNet-18 is the strongest ansatz-OOD predictor on three of four targets: it reduces the mean absolute gap to 0.0066 Ha on EfficientSU2, 0.0048 Ha on HEA-RYRZ-Circular, and 0.0073 Ha on StronglyEntanglingLayers while retaining 11.0–12.7% shot savings. The Transformer remains best on RealAmplitudes at 0.0030 Ha. This result refines the earlier architecture conclusion: token cross-attention is strongest for molecule OOD, whereas convolutional grid inductive bias is unexpectedly robust to unseen ansatz topology. No single predictor dominates every transfer axis.

H.3 Multi-Ansatz Training Improves Transfer

Single-ansatz training isolates a hard transfer question: a selector trained only on HEA-RYRZ-linear circuits can transfer to some named ansatz families, but fails on StronglyEntanglingLayers. Multi-ansatz training keeps the architecture, molecule panel, online VQE protocol, threshold, held-out seeds, and depth-OOD setting fixed, but changes the training distribution. The learner is now trained on 11,520 snapshots from four ansatz families—HEA-RYRZ-linear, HEA-RYRZ-circular, EfficientSU2, and RealAmplitudes—across depths 1–4, then evaluated on depths 5–7.

Table 43: **Multi-ansatz training improves named-ansatz transfer.** The single-ansatz baseline trains the relevance selector only on HEA-RYRZ-linear circuits. The multi-ansatz variant trains the same model architecture on a mixed ansatz panel: HEA-RYRZ-linear, HEA-RYRZ-circular, EfficientSU2, and RealAmplitudes. Both settings evaluate online VQE on the same five STO-3G molecular Hamiltonians, depths 5–7, four held-out seeds per molecule-depth cell, $\tau = 0.997$, and 1024 shots per parameter-shift measurement setting. "Abs-gap gain" is the reduction in mean $|\Delta E|$ from single-ansatz to multi-ansatz training; positive values mean multi-ansatz training improved the final-energy match to Full-QWC.

Target ansatz	Training regime	Mean ΔE	Mean $\lvert\Delta E\rvert \downarrow$	Abs-gap gain ↑	Strict CA seeds	Shot gain ↑	Cells won vs Random
HEA-RYRZ-Circular	Single ansatz	+0.0033	0.0088	–	12/60	12.9%	**15/15**
HEA-RYRZ-Circular	**Ours** multi-ansatz	+0.0015	**0.0032**	**+0.0057**	**34/60**	**14.0%**	**14/15**
EfficientSU2	Single ansatz	+0.0030	0.0119	–	13/60	9.5%	**14/15**
EfficientSU2	**Ours** multi-ansatz	+0.0002	**0.0069**	**+0.0050**	**24/60**	**15.1%**	**15/15**
RealAmplitudes	Single ansatz	−0.0006	**0.0030**	–	**29/60**	13.5%	**15/15**
RealAmplitudes	**Ours** multi-ansatz	−0.0008	0.0034	-0.0004	22/60	**20.8%**	**15/15**
StronglyEntanglingLayers	Single ansatz	+0.0395	**0.0453**	–	0/60	11.1%	**4/15**
StronglyEntanglingLayers	**Ours** multi-ansatz	+0.0360	0.0476	-0.0023	**2/60**	**13.4%**	3/15

The multi-ansatz training distribution substantially improves transfer to HEA-RYRZ-Circular and EfficientSU2: mean absolute gap drops from 0.0088 to 0.0032 Ha for HEA-RYRZ-Circular and from 0.0119 to 0.0069 Ha for EfficientSU2, while strict chemical-accuracy seed counts increase from 12/60 to 34/60 and from 13/60 to 24/60, respectively. RealAmplitudes remains a strong case but shifts toward a more measurement-frugal operating point, increasing shot saving from 13.5% to 20.8% at a small cost in mean absolute gap. StronglyEntanglingLayers remains the hard OOD failure case. This suggests that our method learns transferable structure-conditioned relevance rules when the training panel covers comparable entangler families, but substantially different entangling layers require either explicit gate/topology embeddings or inclusion in the training distribution.

H.4 Controlled Same-Family Ansatz OOD Transfer With Baselines

This comparison extends the controlled same-family ansatz-transfer setting with TITAN-style parameter-freezing baselines under the identical online VQE protocol. For each ansatz family, the learned selector is trained on two topology variants and evaluated on held-out same-family topology variants. The test ansatz is unseen during training, but the transfer is controlled: same rotation family, same molecular panel, train depths 1–4, unseen evaluation depths 5–7, four held-out seeds per molecule-depth cell, $\tau = 0.997$, and 1024 shots per parameter-shift measurement setting.

Table 44: **Controlled same-family ansatz OOD transfer with matched baselines at $\tau = 0.997$.** Each block trains our method on two topology variants from one ansatz family and evaluates on held-out same-family topologies that do not appear in the training labels. TITAN (Peng et al., 2026) applies active-parameter freezing and full-QWC measurements for active parameters; TITAN+**Ours** composes the same active-parameter mask with our method group selection. CA excess reports $1000 \max(\overline{\Delta E} - 0.0016, 0)$ in mHa, so lower energies are accepted and smaller is better. "Cells won" counts molecule-depth cells, out of 15, where the method has smaller mean $|\Delta E|$ than matched-budget Random.

Family / train topologies	Held-out target	Method	Mean ΔE	Mean $\lvert\Delta E\rvert$ ↓	Shot gain ↑	CA excess (mHa) ↓	Cells won
HEA-RYRZ / linear + reverse-linear	Circular	**Ours**	**+0.0017**	**0.0055**	12.40%	**0.09**	**14/15**
HEA-RYRZ / linear + reverse-linear	Circular	Random	+0.0614	0.0716	12.70%	59.80	–
HEA-RYRZ / linear + reverse-linear	Circular	TITAN (Peng et al.)	+0.0074	**0.0458**	33.00%	5.77	6/15
HEA-RYRZ / linear + reverse-linear	Circular	TITAN+**Ours**	+0.0074	**0.0451**	**40.10%**	5.82	6/15
HEA-RYRZ / linear + reverse-linear	Full	**Ours**	**-0.0002**	**0.0085**	8.50%	**0.00**	**14/15**
HEA-RYRZ / linear + reverse-linear	Full	Random	+0.0280	0.0460	8.50%	26.45	–
HEA-RYRZ / linear + reverse-linear	Full	TITAN (Peng et al.)	+0.0484	0.1007	40.80%	46.82	2/15
HEA-RYRZ / linear + reverse-linear	Full	TITAN+**Ours**	+0.0481	0.1010	**44.60%**	46.54	2/15
EfficientSU2 / linear + reverse-linear	Circular	**Ours**	**+0.0019**	**0.0056**	12.40%	**0.27**	**15/15**
EfficientSU2 / linear + reverse-linear	Circular	Random	+0.0612	0.0716	12.60%	59.59	–
EfficientSU2 / linear + reverse-linear	Circular	TITAN (Peng et al.)	+0.0074	**0.0458**	33.00%	5.77	6/15
EfficientSU2 / linear + reverse-linear	Circular	TITAN+**Ours**	+0.0073	**0.0450**	**40.10%**	5.69	6/15
EfficientSU2 / linear + reverse-linear	Full	**Ours**	**-0.0002**	**0.0089**	8.60%	**0.00**	**12/15**
EfficientSU2 / linear + reverse-linear	Full	Random	+0.0377	0.0648	8.50%	36.11	–
EfficientSU2 / linear + reverse-linear	Full	TITAN (Peng et al.)	+0.0484	0.1007	40.80%	46.82	3/15
EfficientSU2 / linear + reverse-linear	Full	TITAN+**Ours**	+0.0481	0.1009	**44.60%**	46.51	3/15
RealAmplitudes / linear + reverse-linear	Circular	**Ours**	**+0.0052**	**0.0177**	19.60%	3.63	**13/15**
RealAmplitudes / linear + reverse-linear	Circular	Random	+0.0684	0.0858	19.90%	66.75	–
RealAmplitudes / linear + reverse-linear	Circular	TITAN (Peng et al.)	**-0.0245**	**0.0512**	18.00%	**0.00**	11/15
RealAmplitudes / linear + reverse-linear	Circular	TITAN+**Ours**	**-0.0240**	**0.0527**	**33.40%**	**0.00**	10/15
RealAmplitudes / linear + reverse-linear	Full	**Ours**	**+0.0012**	**0.0235**	13.90%	**0.00**	**14/15**
RealAmplitudes / linear + reverse-linear	Full	Random	+0.0648	0.0880	14.00%	63.18	–
RealAmplitudes / linear + reverse-linear	Full	TITAN (Peng et al.)	**-0.0014**	0.0885	22.80%	**0.00**	9/15
RealAmplitudes / linear + reverse-linear	Full	TITAN+**Ours**	**-0.0014**	0.0899	**33.90%**	**0.00**	9/15
StronglyEntangling / range-1 + range-2	Layer-dependent	**Ours**	**+0.0012**	**0.0077**	12.40%	**0.00**	**13/15**
StronglyEntangling / range-1 + range-2	Layer-dependent	Random	+0.0343	0.0447	12.60%	32.69	–
StronglyEntangling / range-1 + range-2	Layer-dependent	TITAN (Peng et al.)	+0.0104	0.0454	39.60%	8.80	4/15
StronglyEntangling / range-1 + range-2	Layer-dependent	TITAN+**Ours**	+0.0106	0.0456	**45.50%**	9.01	5/15
StronglyEntangling / range-1 + range-2	Full	**Ours**	**+0.0023**	**0.0206**	8.90%	**0.66**	**13/15**
StronglyEntangling / range-1 + range-2	Full	Random	+0.0596	0.0968	9.10%	57.97	–
StronglyEntangling / range-1 + range-2	Full	TITAN (Peng et al.)	+0.0493	0.1071	47.30%	47.68	4/15
StronglyEntangling / range-1 + range-2	Full	TITAN+**Ours**	+0.0502	0.1066	**50.70%**	48.59	4/15

Table 44 shows that the controlled same-family transfer result is not explained by measurement budget alone. Matched-budget Random keeps similar shot savings but is substantially less accurate. TITAN-style freezing can reduce cost by dropping parameters, and TITAN+**Ours** can compound this saving, but these rows often trade away final-energy fidelity. The cleanest read is that our method targets a different axis from parameter freezing: for parameters that are still measured, it learns which QWC groups are worth measuring.

H.5 CELL-WISE BREAKDOWN OF TABLE 44

Table 45 expands the "Cells won" column from Table 44. Each entry reports the molecule-depth cell improvement in mean absolute final-energy gap, measured in mHa:

$$1000 \times \left(|\Delta E|_{\text{Random}} - |\Delta E|_{\text{Ours}}\right). \tag{9}$$

Positive green values mean Ours is closer to Full-QWC for that exact molecule-depth cell; negative red values indicate a random same-budget win.

Table 45: **Cell-wise breakdown of the same-family ansatz-OOD wins.** Entries are Random-minus-Ours mean $|\Delta E|$ in mHa for each molecule and depth. Green positive entries are cells won by Ours; red negative entries are cells won by the matched-budget random baseline.

Held-out target	H2			HeH+			H3+			H4			LiH			Wins
	D5	D6	D7	D5	D6	D7	D5	D6	D7	D5	D6	D7	D5	D6	D7	
HEA-RYRZ / Circular	**+11.4**	**+4.8**	-0.4	**+30.9**	**+5.0**	**+3.0**	**+18.0**	**+16.0**	**+8.7**	**+22.7**	**+47.7**	**+36.8**	**+320.2**	**+198.4**	**+268.8**	**14/15**
HEA-RYRZ / Full	**+18.1**	**+22.6**	-0.0	**+37.0**	**+10.2**	**+10.8**	**+46.5**	**+32.6**	**+0.7**	**+22.9**	**+8.3**	**+51.8**	**+42.9**	**+109.2**	**+148.2**	**14/15**
EfficientSU2 / Circular	**+3.5**	**+5.3**	**+0.3**	**+28.0**	**+5.7**	**+0.9**	**+14.5**	**+13.3**	**+16.6**	**+25.2**	**+53.8**	**+111.6**	**+280.9**	**+215.2**	**+214.5**	**15/15**
EfficientSU2 / Full	**+18.1**	**+14.8**	-1.8	**+31.5**	-4.5	**+22.9**	**+38.6**	**+13.8**	-0.2	**+125.8**	**+20.9**	**+51.5**	**+109.9**	**+206.7**	**+190.7**	**12/15**
RealAmplitudes / Circular	**+19.9**	**+7.8**	-2.1	**+94.3**	**+0.5**	**+42.1**	**+24.5**	**+18.2**	-8.1	**+80.6**	**+33.9**	**+63.8**	**+188.2**	**+190.3**	**+267.0**	**13/15**
RealAmplitudes / Full	**+34.1**	-1.6	**+0.0**	**+33.1**	**+27.2**	**+99.1**	**+48.6**	**+121.0**	**+39.5**	**+121.7**	**+18.1**	**+5.6**	**+148.9**	**+75.9**	**+195.2**	**14/15**
StronglyEntangling / Layer-dependent	-3.6	**+12.9**	**+2.7**	**+15.4**	**+18.1**	**+4.9**	**+13.7**	-1.7	**+27.0**	**+39.2**	**+2.6**	**+29.1**	**+183.9**	**+84.9**	**+125.5**	**13/15**
StronglyEntangling / Full	**+10.7**	**+9.0**	**+21.3**	**+83.2**	-14.9	**+11.3**	**+4.9**	**+14.1**	**+20.1**	**+62.3**	-23.3	**+5.3**	**+473.9**	**+249.6**	**+215.9**	**13/15**

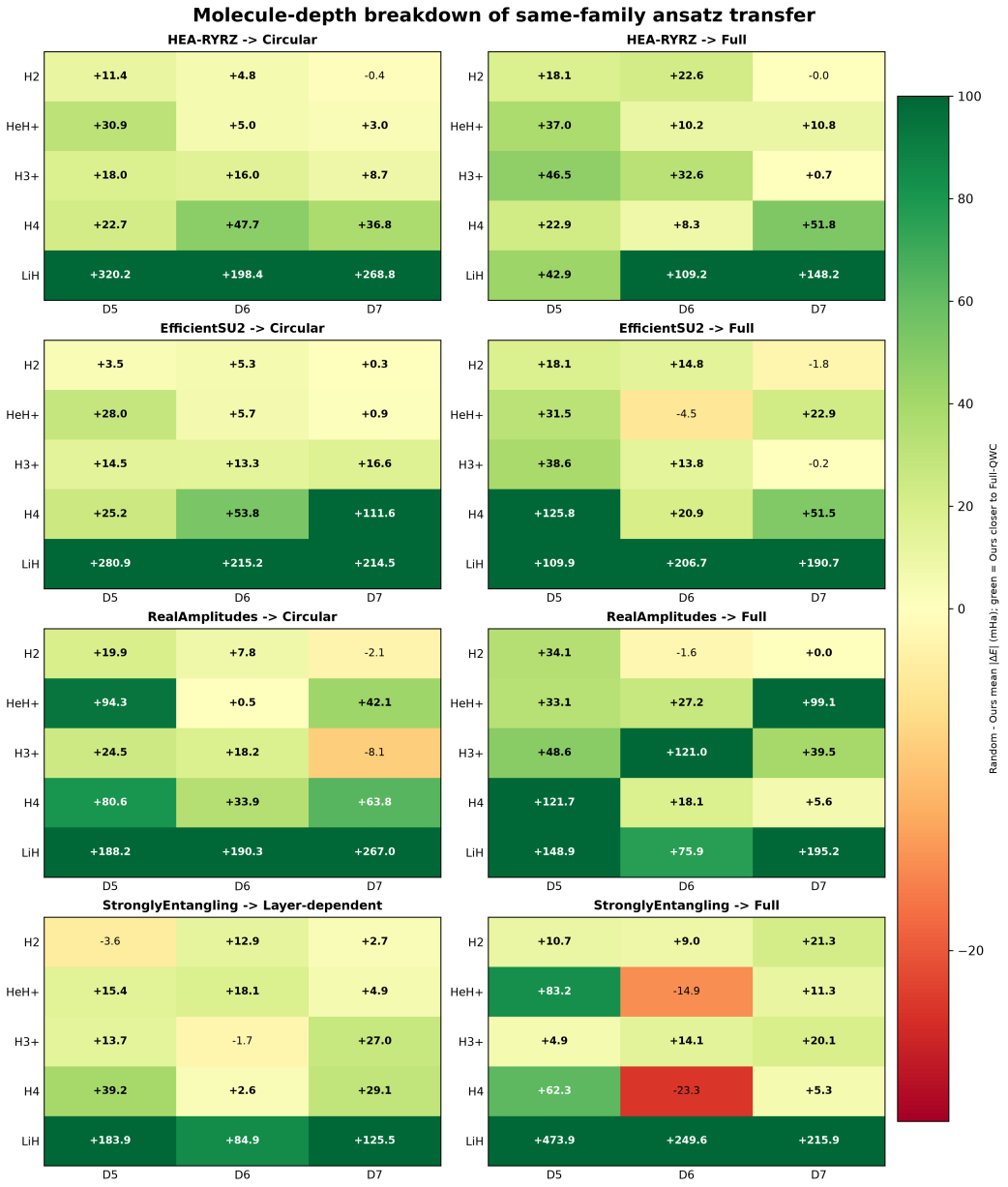

Figure 9: **Molecule-depth breakdown for Table 44.** Each panel corresponds to one controlled same-family ansatz-OOD held-out target. Rows are molecules, columns are unseen evaluation depths 5–7, and entries are Random-minus-**Ours** mean $|\Delta E|$ in mHa. Green positive entries are cells where our method is closer to Full-QWC than matched-budget Random. The color scale is clipped for readability, but numeric labels show the unclipped values rounded to one decimal place.

H.6 Molecule-Depth Breakdown of Table 44

Figure 9 expands the "Cells won" column from Table 44 into individual molecule-depth cells. Each entry reports

$$1000 \times \left(|\Delta E|_{\text{Random}} - |\Delta E|_{\text{Ours}}\right), \tag{10}$$

measured in mHa. Positive green values mean our method is closer to Full-QWC for that exact molecule-depth cell; negative red values mean the matched-budget random selector is closer.

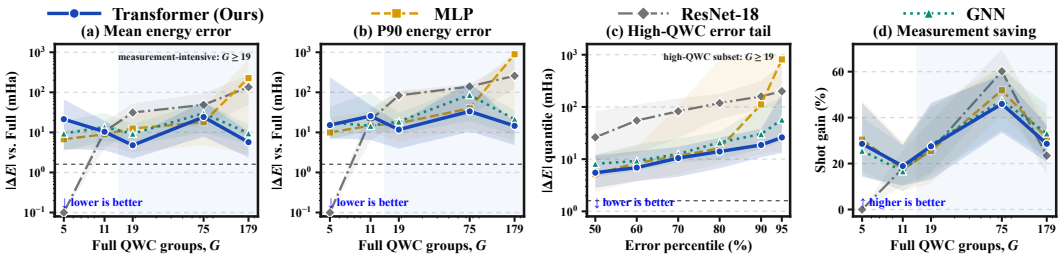

Figure 10: **Predictor robustness as QWC measurement complexity grows.** We use leave-one-molecule-out transfer at $\tau = 0.997$ and order the five held-out molecular Hamiltonians by their Full-QWC group counts $G \in \{5, 11, 19, 75, 179\}$. Panels (a,b) report the mean and 90th-percentile absolute final-energy gap to Full QWC; panel (c) shows the error-quantile curve over the three measurement-intensive targets with $G \geq 19$; and panel (d) reports mean shot gain. Lines are computed from four depths and four paired seeds, and bands are hierarchical-bootstrap 95% confidence intervals. The Transformer has the lowest P90 error at all three measurement-intensive group counts while retaining a measurement budget comparable to the other learned predictors. The dashed line marks 1.6 mHa. All features, labels, and online-VQE settings are fixed while only the predictor architecture changes. A cell-level mask diagnostic for the same predictors appears in Appendix I, Figure 11; underlying molecule-OOD values and matched held-out-seed/Hamiltonian controls appear in Tables 47, 46, and 48.

I Predictor Architecture Tables

Figure 10 gives the aggregate architecture comparison referenced in the main text. Figure 11 provides its cell-level counterpart. On this held-out snapshot, the Transformer captures 96.8% of the exact relevance mass while selecting 76.8% of parameter–group cells. ResNet-18 captures slightly more mass (98.6%) by selecting 89.3% of cells, whereas the GNN is sparser but misses more oracle-relevant groups. Similar scalar budgets can therefore arise from materially different parameter-specific masks.

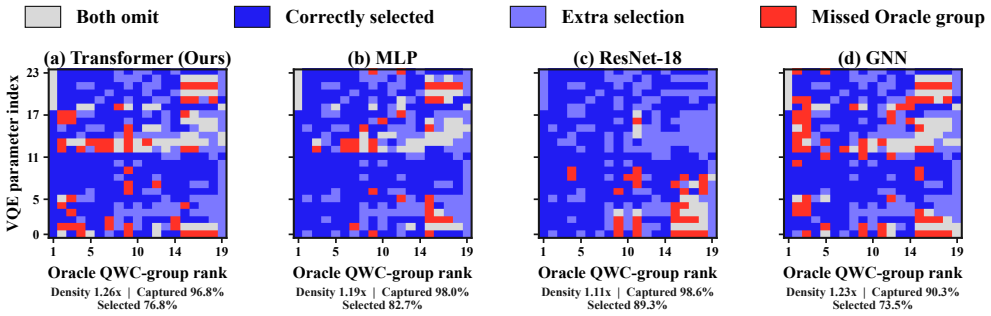

Figure 11: **Oracle-relative parameter–QWC masks for four relevance predictors.** All predictors are evaluated on the same held-out H_3^+ molecule-OOD snapshot (depth 2, seed 97, optimization iteration 14) at $\tau = 0.997$. Rows index VQE parameters and columns index QWC groups sorted by oracle relevance. Dark-blue cells are selected by both the predictor and oracle; gray cells are omitted by both; light-blue cells are predictor-only extra measurements; and red cells are oracle-selected groups missed by the predictor. "Captured" is the mean exact relevance mass covered by a predicted row mask, "Selected" is its mean selected-group fraction, and "Density" is captured mass divided by selected fraction. The snapshot is a qualitative structural diagnostic; aggregate performance is reported in main-text Figure 10 and Tables 47 and 48.

Table 46: **Predictor architecture comparison at $\tau = 0.995$.** This table copies the Table 1 held-out-seed online VQE setting and changes only the learned relevance predictor. Our method uses the Transformer predictor throughout the paper; MLP, ResNet-18, and GNN use the same labels, features, threshold, molecules, depths, and online protocol. Boldface marks the closest final energy; shot-gain values are green/red according to the no-worse criterion, and bold-underlined values mark the largest shot gain per cell.

Setting	Metric	Method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 1	ΔE vs Full ↓	**Ours**	+0.0011	+0.0015	+0.0005	−0.0006	+0.0023
Depth 1	ΔE vs Full ↓	MLP	+0.0008	+0.0028	+0.0015	−0.0021	+0.0025
Depth 1	ΔE vs Full ↓	ResNet-18	+0.0000	−0.0010	+0.0029	+0.0024	+0.0016
Depth 1	ΔE vs Full ↓	GNN	+0.0012	+0.0022	+0.0004	−0.0021	+0.0013
Depth 1	Shot gain ↑	**Ours**	40.4%	58.9%	63.0%	74.7%	86.9%
Depth 1	Shot gain ↑	MLP	**47.3%**	**62.5%**	**68.2%**	**75.9%**	**87.6%**
Depth 1	Shot gain ↑	ResNet-18	0.0%	7.5%	38.1%	46.1%	77.9%
Depth 1	Shot gain ↑	GNN	42.7%	57.7%	60.3%	66.6%	82.5%
Depth 1	No-worse vs Full	**Ours**	✓	✓	✓	✓	✗
Depth 1	No-worse vs Full	MLP	✓	✗	✓	✓	✗
Depth 1	No-worse vs Full	ResNet-18	✓	✓	✗	✗	✓
Depth 1	No-worse vs Full	GNN	✓	✗	✓	✓	✓
Depth 2	ΔE vs Full ↓	**Ours**	+0.0003	−0.0011	−0.0015	+0.0158	−0.0049
Depth 2	ΔE vs Full ↓	MLP	+0.0127	−0.0014	−0.0029	+0.0006	−0.0037
Depth 2	ΔE vs Full ↓	ResNet-18	+0.0000	+0.0011	+0.0007	+0.0221	+0.0065
Depth 2	ΔE vs Full ↓	GNN	+0.0010	−0.0134	+0.0031	+0.0254	−0.0026
Depth 2	Shot gain ↑	**Ours**	25.6%	31.6%	44.9%	**51.8%**	**82.6%**
Depth 2	Shot gain ↑	MLP	**28.3%**	**33.4%**	**46.1%**	48.4%	78.0%
Depth 2	Shot gain ↑	ResNet-18	0.0%	2.2%	24.2%	35.2%	73.3%
Depth 2	Shot gain ↑	GNN	21.6%	28.3%	41.7%	45.9%	77.6%
Depth 2	No-worse vs Full	**Ours**	✓	✓	✓	✗	✓
Depth 2	No-worse vs Full	MLP	✗	✓	✓	✓	✓
Depth 2	No-worse vs Full	ResNet-18	✓	✓	✓	✗	✗
Depth 2	No-worse vs Full	GNN	✓	✓	✗	✗	✓
Depth 3	ΔE vs Full ↓	**Ours**	+0.0005	+0.0092	+0.0005	−0.0014	−0.0017
Depth 3	ΔE vs Full ↓	MLP	+0.0035	−0.0045	−0.0008	−0.0000	−0.0039
Depth 3	ΔE vs Full ↓	ResNet-18	+0.0000	+0.0044	−0.0004	−0.0064	+0.0053
Depth 3	ΔE vs Full ↓	GNN	+0.0039	−0.0017	+0.0020	−0.0025	+0.0001
Depth 3	Shot gain ↑	**Ours**	18.1%	**24.9%**	**29.9%**	**35.2%**	**75.4%**
Depth 3	Shot gain ↑	MLP	**20.8%**	23.7%	28.9%	32.7%	71.5%
Depth 3	Shot gain ↑	ResNet-18	0.0%	2.0%	19.9%	28.4%	73.0%
Depth 3	Shot gain ↑	GNN	16.8%	22.1%	28.4%	31.7%	66.9%
Depth 3	No-worse vs Full	**Ours**	✓	✗	✓	✓	✓
Depth 3	No-worse vs Full	MLP	✗	✓	✓	✓	✓
Depth 3	No-worse vs Full	ResNet-18	✓	✗	✓	✓	✗
Depth 3	No-worse vs Full	GNN	✗	✓	✗	✓	✓

I.1 Held-Out-Seed Predictor Comparison

I.2 Leave-One-Molecule-Out Predictor Comparison

Table 47: **Predictor comparison under leave-one-molecule-out transfer at $\tau = 0.997$.** The predictor is the only changed component: all methods use identical features, labels, molecule splits, depths, seeds, and closed-loop online VQE. Representative depths 1 and 4 are shown. Green/red follows the one-sided 1.6 mHa no-worse criterion; bold marks the closest mean energy to Full, and bold-underlined shot gain marks the largest saving per cell.

Molecule	Depth	Metric	**Ours**	MLP	ResNet-18	GNN
H2 (4q)	Depth 1	ΔE vs Full ↓	+0.0016	+0.0030	**+0.0000**	+0.0018
	Depth 1	Shot gain ↑	55.2%	**56.7%**	0.0%	46.9%
	Depth 1	Chem. acc.	✗	✗	✓	✗
	Depth 4	ΔE vs Full ↓	**-0.0002**	**-0.0020**	**+0.0000**	**-0.0000**
	Depth 4	Shot gain ↑	9.9%	**13.7%**	0.0%	12.2%
	Depth 4	Chem. acc.	✓	✓	✓	✓
HeH+ (4q)	Depth 1	ΔE vs Full ↓	**+0.0021**	**-0.0038**	+0.0139	**-0.0032**
	Depth 1	Shot gain ↑	32.2%	**36.2%**	27.6%	33.6%
	Depth 1	Chem. acc.	✗	✓	✗	✓
	Depth 4	ΔE vs Full ↓	**-0.0022**	**-0.0025**	**-0.0071**	**-0.0018**
	Depth 4	Shot gain ↑	8.1%	6.0%	**9.8%**	6.0%
	Depth 4	Chem. acc.	✓	✓	✓	✓
H3+ (6q)	Depth 1	ΔE vs Full ↓	+0.0051	**-0.0004**	+0.0078	+0.0034
	Depth 1	Shot gain ↑	**56.2%**	55.2%	45.3%	52.3%
	Depth 1	Chem. acc.	✗	✓	✗	✗
	Depth 4	ΔE vs Full ↓	**-0.0052**	**-0.0057**	**-0.0320**	**+0.0029**
	Depth 4	Shot gain ↑	10.0%	8.9%	**16.1%**	12.3%
	Depth 4	Chem. acc.	✓	✓	✓	✗
H4 chain (8q)	Depth 1	ΔE vs Full ↓	+0.0038	+0.0052	**+0.0030**	+0.0072
	Depth 1	Shot gain ↑	63.7%	**67.1%**	59.8%	66.3%
	Depth 1	Chem. acc.	✗	✗	✗	✗
	Depth 4	ΔE vs Full ↓	+0.0652	+0.0195	+0.0335	**-0.0043**
	Depth 4	Shot gain ↑	28.5%	35.7%	**46.5%**	28.9%
	Depth 4	Chem. acc.	✗	✗	✗	✓
LiH (12q)	Depth 1	ΔE vs Full ↓	**+0.0022**	+0.8939	**-0.0314**	+0.0101
	Depth 1	Shot gain ↑	40.4%	47.1%	41.4%	**51.9%**
	Depth 1	Chem. acc.	✗	✗	✓	✗
	Depth 4	ΔE vs Full ↓	**-0.0023**	+0.0037	+0.1054	+0.0073
	Depth 4	Shot gain ↑	17.9%	19.4%	13.5%	**19.5%**
	Depth 4	Chem. acc.	✓	✗	✗	✗

I.3 Hamiltonian-Family Predictor Comparison

Table 48: **Predictor comparison under Hamiltonian-family transfer at $\tau = 0.997$.** One shared predictor per architecture is trained on the same five molecular and two Heisenberg Hamiltonians, then evaluated on seven held-out Hamiltonians with identical features, depths, seeds, and online VQE. Representative depths 1 and 4 are shown. A zero-gap ResNet-18 cell with zero shot gain denotes Full-QWC fallback rather than successful sparsification.

Testing Hamiltonian	Depth	Metric	**Ours**	MLP	ResNet-18	GNN
TFIM $h = 0.5$	Depth 1	ΔE vs Full ↓	**-0.0000**	**-0.0000**	**+0.0000**	**+0.0000**
	Depth 1	Shot gain ↑	2.0%	**5.7%**	0.0%	2.1%
	Depth 1	Chem. acc.	✓	✓	✓	✓
	Depth 4	ΔE vs Full ↓	**-0.0002**	**+0.0000**	**+0.0000**	**+0.0000**
	Depth 4	Shot gain ↑	2.6%	**4.5%**	0.0%	1.3%
	Depth 4	Chem. acc.	✓	✓	✓	✓
TFIM $h = 0.7$	Depth 1	ΔE vs Full ↓	**-0.0000**	**-0.0000**	**+0.0000**	**+0.0000**
	Depth 1	Shot gain ↑	2.9%	**4.9%**	0.0%	1.7%
	Depth 1	Chem. acc.	✓	✓	✓	✓
	Depth 4	ΔE vs Full ↓	**+0.0000**	**-0.0000**	**+0.0000**	**-0.0014**
	Depth 4	Shot gain ↑	2.4%	**4.3%**	0.0%	1.2%
	Depth 4	Chem. acc.	✓	✓	✓	✓
TFIM $h = 0.9$	Depth 1	ΔE vs Full ↓	+0.0771	+0.0771	**+0.0000**	+0.0496
	Depth 1	Shot gain ↑	**25.0%**	**25.0%**	0.0%	14.1%
	Depth 1	Chem. acc.	✗	✗	✓	✗
	Depth 4	ΔE vs Full ↓	+0.0302	+0.0302	**+0.0000**	+0.0255
	Depth 4	Shot gain ↑	**6.2%**	**6.2%**	0.0%	3.8%
	Depth 4	Chem. acc.	✗	✗	✓	✗
Heis. XXZ+X field	Depth 1	ΔE vs Full ↓	**+0.0000**	**+0.0000**	**+0.0000**	**+0.0000**
	Depth 1	Shot gain ↑	**16.7%**	**16.7%**	0.0%	16.4%
	Depth 1	Chem. acc.	✓	✓	✓	✓
	Depth 4	ΔE vs Full ↓	**+0.0000**	**+0.0000**	**+0.0000**	**+0.0000**
	Depth 4	Shot gain ↑	**4.2%**	**4.2%**	0.0%	**4.2%**
	Depth 4	Chem. acc.	✓	✓	✓	✓
Heis. J1–J2	Depth 1	ΔE vs Full ↓	**-0.0000**	**+0.0000**	**+0.0000**	**+0.0001**
	Depth 1	Shot gain ↑	16.7%	**16.8%**	0.0%	15.9%
	Depth 1	Chem. acc.	✓	✓	✓	✓
	Depth 4	ΔE vs Full ↓	**+0.0000**	**+0.0000**	**+0.0000**	**+0.0001**
	Depth 4	Shot gain ↑	4.2%	4.2%	0.0%	**4.2%**
	Depth 4	Chem. acc.	✓	✓	✓	✓
Heis. XXY+Z field	Depth 1	ΔE vs Full ↓	**+0.0000**	**+0.0001**	**+0.0000**	**+0.0000**
	Depth 1	Shot gain ↑	16.7%	**16.7%**	0.0%	15.8%
	Depth 1	Chem. acc.	✓	✓	✓	✓
	Depth 4	ΔE vs Full ↓	**+0.0000**	**+0.0000**	**+0.0000**	**+0.0000**
	Depth 4	Shot gain ↑	**4.2%**	**4.2%**	0.0%	**4.2%**
	Depth 4	Chem. acc.	✓	✓	✓	✓
JW-like parity spin	Depth 1	ΔE vs Full ↓	+0.0312	**+0.0290**	+0.0452	+0.0337
	Depth 1	Shot gain ↑	35.7%	23.0%	**65.8%**	31.9%
	Depth 1	Chem. acc.	✗	✗	✗	✗
	Depth 4	ΔE vs Full ↓	**-0.0046**	**-0.0014**	+0.1674	+0.0049
	Depth 4	Shot gain ↑	11.6%	8.8%	**28.2%**	15.1%
	Depth 4	Chem. acc.	✓	✓	✗	✗

J FROZEN-LLM AND FRONTIER-MODEL DETAILS

Controlled intervention. The local-LLM experiment does not replace our full controller and does not fine-tune an LLM. Instead, it isolates one decision: ranking QWC groups for each parameter. At a refresh iteration t, our saved controller first predicts a binary mask and therefore a per-parameter cardinality

$$k_{j,t} = \sum_{g=1}^{G} M_{j,g}^{\mathrm{Ours}}(t). \tag{11}$$

The LLM receives $k_{j,t}$ and the same structured parameter, group, and parameter–group features, but it never receives the oracle contribution matrix. It returns a ranking $r_{j,t}^{\mathrm{LLM}}$, from which we form

$$M_{j,g}^{\mathrm{LLM}}(t) = \mathbb{I}\Big[g \in \mathrm{Top}_{k_{j,t}}\big(r_{j,t}^{\mathrm{LLM}}\big)\Big]. \tag{12}$$

Thus the comparison holds the count controller fixed at each query and tests whether a frozen language model can replace the learned group-ranking function.

Exact system prompt for the repaired rank-only protocol. All four models receive the following system message without demonstrations:

```
You are a measurement-selection controller for variational
quantum eigensolvers.
For each requested circuit parameter, choose exactly k
distinct QWC group IDs and order them from
most to least relevant to that parameter's energy gradient.
Use coefficient, circuit-location,
theta, iteration, and parameter--group light-cone features
jointly.  A globally large coefficient
does not imply equal relevance to every parameter.  Every
returned ID must be an integer present in
the candidate groups.  Return only the JSON object required
by the supplied schema.
```

Compact request and response. The request contains only feature columns that are nonzero in the current instance. The ellipses below suppress repeated rows, not additional natural-language instructions.

```
{
  "task": "rank_parameter_specific_qwc_groups",
  "context": {
    "ansatz": "ryrz", "depth": 2,
    "parameters": 48, "qwc_groups": 19,
    "iteration": 7, "total_iterations": 28
  },
  "columns": {
    "parameter_x": [...],
    "group_rows": ["group_id", ...],
    "pair_nonzero_rows": ["group_id", ...]
  },
  "groups": [[0, ...], [1, ...], ...],
  "parameters_to_rank": [
    {"id": 3, "k": 5, "x": [...],
     "pair_nonzero": [[0, ...], ...]}
  ]
}
```

The response is constrained by Ollama's JSON schema and contains one row for every requested parameter:

```
{"selections": [
```

```
  {"parameter_id": 3,
   "ranked_group_ids": [7, 2, 5, 1, 8]}
]}
```

Validation and repair in the rank-only protocol. Later clean/explanation output contracts are different; Appendix N.4 documents the clean and mechanism-audit protocols and their prompt-specific trace records. Boolean values, repeated IDs, noninteger values, and group IDs outside $[0, G-1]$ are discarded. If fewer than $k_{j,t}$ valid distinct IDs remain, the missing positions are filled deterministically using descending global group-coefficient magnitude. We log the requested and repaired group counts, invalid batches, context splits, prompt/output tokens, and inference latency. This repair prevents malformed JSON from changing the requested cardinality while preserving a measurable repair-rate diagnostic.

Sparse-refresh execution. The LLM is queried only at $t \in \{0, 7, 14, 21\}$ during each 28-step online trajectory. The resulting mask is reused for the following seven optimizer steps. Because the count targets predicted by our method can evolve between refreshes while the LLM mask is held fixed, an LLM's realized cumulative shot gain need not equal the continuously updated reference gain from our method. We therefore treat the reference-gain rows in Table 3 as count-controller context rather than as measured LLM cost.

Table 49: Frozen local-LLM inference settings used in the sparse-refresh comparison.

Component	Setting
Models	Qwen3-8B, Llama-3.1-8B, DeepSeek-R1-8B, GPT-OSS-20B
Adaptation	Zero-shot; no demonstrations; no fine-tuning
Decoding	Temperature 0, seed 123, `think=false`
Context/output	`num_ctx`= 32768, `num_predict`= 1536
Parameter batching	Up to 16 parameters; adaptive halving on context overflow
Refresh schedule	Iterations $\{0, 7, 14, 21\}$ of 28
VQE setting	1024 shots per shift, Adam, layered RY/RZ ansatz
Evaluation panel	Five STO-3G molecules, depths 1–3, seeds 97/101/107/109
Count controller	Saved **Ours** model at $\tau = 0.995$

Controlled rank-only intervention. We retain our method only as the cardinality controller: at each query it supplies k_j, while a frozen local LLM must infer *which* QWC groups to select from structured parameter, group, and light-cone features. The oracle contribution matrix is never included in the prompt.

No model adaptation. Qwen3-8B, Llama-3.1-8B, DeepSeek-R1-8B, and GPT-OSS-20B are evaluated zero-shot with deterministic decoding and no fine-tuning. Schema validation and coefficient-based repair prevent malformed output from changing the requested cardinality.

Sparse refresh. The LLM is queried at $t \in \{0, 7, 14, 21\}$ and each returned mask is reused for seven online-VQE steps. The complete prompt and settings are in Appendix J.

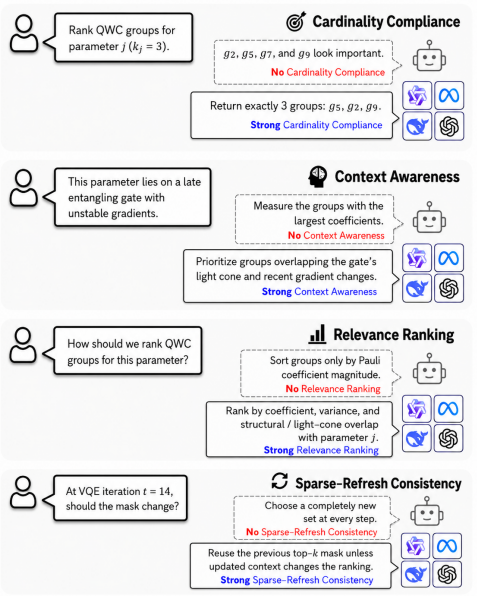

Figure 12: **Controlled diagnostics for frozen local LLMs.** The schematic contrasts failure modes (red) and desired behavior (blue) for cardinality compliance, context awareness, parameter-specific QWC relevance ranking, and sparse-refresh consistency.

Table 50: **Clean frozen local-LLM QWC-selection audit.** This table duplicates the molecule–depth protocol and layout of Table 3, but replaces its repaired rank-only outputs with the stricter exact select/exclude contract. Each frozen LLM must emit unique QWC IDs at the requested per-parameter cardinality; invalid responses receive schema feedback for at most two retries, with no coefficient-based repair. Persistent failures use a logged Full-QWC safety fallback and pay the corresponding shots. Random uses the same per-parameter cardinality predicted by our method and $K=7$ held-mask cadence, but samples group identities uniformly. All four models complete 60 online trajectories over seeds 97/101/107/109; their aggregate realized shot savings lie between 52.0% and 52.4%, versus 52.7% for Random. Energy entries are mean ± standard deviation in Ha relative to Full-QWC, using exact-expectation gradients and nominal shot accounting (Appendix N.1). Random cells are intentionally unshaded. For every other method, green/red satisfies/violates the one-sided 1.6 mHa no-worse criterion ($\Delta E \leq 1.6$ mHa; lower energy is accepted); deeper color indicates a lower energy/larger violation. The black check/cross separately indicates whether the absolute mean ΔE is smaller/not smaller than Random. Bold/underline mark the smallest/second-smallest absolute mean ΔE per cell. Figure 4 summarizes the audited protocol.

Setting	Selection method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 1	**Ours** gain, $K=1\uparrow$	40.4 ± 6.3%	58.9 ± 3.5%	63.0 ± 2.8%	74.7 ± 1.1%	86.9 ± 0.5%
Depth 1	**Ours** gain, $K=7\uparrow$	41.2 ± 5.3%	60.7 ± 3.1%	65.1 ± 2.6%	77.3 ± 0.7%	88.2 ± 0.4%
Depth 1	Random ($K=7$)	+0.1847 ± 0.0868	+1.2191 ± 0.2759	+0.7247 ± 0.2856	+0.9736 ± 0.3101	+2.9030 ± 0.6058
Depth 1	Qwen3-8B	+0.2012 ± 0.1085 ✗	+0.9101 ± 0.4857 ✓	+0.0200 ± 0.0203 ✓	+0.9166 ± 0.5180 ✓	+1.6190 ± 0.9510 ✓
Depth 1	Llama-3.1-8B	+0.3306 ± 0.0615 ✗	+0.8140 ± 0.2673 ✓	+0.5751 ± 0.2637 ✓	+0.4804 ± 0.3100 ✓	+0.6219 ± 0.5640 ✓
Depth 1	DeepSeek-R1-8B	+0.2245 ± 0.0696 ✗	+1.0313 ± 0.3889 ✓	+0.5259 ± 0.0514 ✓	+0.6296 ± 0.1083 ✓	+1.2793 ± 0.7197 ✓
Depth 1	GPT-OSS-20B	+0.3930 ± 0.1936 ✗	+0.8238 ± 0.2303 ✓	+0.7300 ± 0.4383 ✗	+0.5671 ± 0.3164 ✓	+1.5503 ± 0.7644 ✓
Depth 1	**Ours** ($K=7$)	+0.0013 ± 0.0003 ✓	+0.0035 ± 0.0015 ✓	**+0.0004 ± 0.0006** ✓	**+0.0001 ± 0.0060** ✓	+0.0030 ± 0.0016 ✓
Depth 1	**Ours** ($K=1$)	**+0.0011 ± 0.0008** ✓	**+0.0015 ± 0.0012** ✓	+0.0005 ± 0.0006 ✓	−0.0006 ± 0.0024 ✓	**+0.0023 ± 0.0016** ✓
Depth 2	**Ours** gain, $K=1\uparrow$	25.6 ± 2.9%	31.6 ± 2.6%	44.9 ± 1.3%	51.8 ± 0.5%	82.6 ± 0.4%
Depth 2	**Ours** gain, $K=7\uparrow$	27.3 ± 2.5%	32.5 ± 2.4%	47.9 ± 1.4%	55.1 ± 0.5%	83.7 ± 0.4%
Depth 2	Random ($K=7$)	+0.0930 ± 0.0297	+0.5668 ± 0.4463	+0.3054 ± 0.1633	+0.9327 ± 0.3273	+2.6969 ± 0.2816
Depth 2	Qwen3-8B	+0.0861 ± 0.0637 ✓	+0.5614 ± 0.3196 ✓	+0.0206 ± 0.0054 ✓	+0.3825 ± 0.1254 ✓	+2.7674 ± 0.5685 ✗
Depth 2	Llama-3.1-8B	+0.0184 ± 0.2098 ✓	+0.8950 ± 0.1642 ✗	+0.5715 ± 0.0934 ✗	+0.2758 ± 0.1151 ✓	+0.7222 ± 0.3456 ✓
Depth 2	DeepSeek-R1-8B	+0.1366 ± 0.1334 ✗	+0.3298 ± 0.3566 ✓	+0.5573 ± 0.1298 ✗	+0.7876 ± 0.2497 ✓	+2.4135 ± 0.3556 ✓
Depth 2	GPT-OSS-20B	+0.2148 ± 0.1220 ✗	+0.6807 ± 0.7081 ✗	+0.4691 ± 0.2024 ✗	+0.5616 ± 0.2675 ✓	+1.2894 ± 0.3290 ✓
Depth 2	**Ours** ($K=7$)	−0.0007 ± 0.0015 ✓	+0.0076 ± 0.0176 ✓	**+0.0006 ± 0.0031** ✓	+0.0162 ± 0.0438 ✓	**−0.0040 ± 0.0118** ✓
Depth 2	**Ours** ($K=1$)	**+0.0003 ± 0.0007** ✓	**−0.0011 ± 0.0026** ✓	−0.0015 ± 0.0041 ✓	**+0.0158 ± 0.0443** ✓	−0.0049 ± 0.0097 ✓
Depth 3	**Ours** gain, $K=1\uparrow$	18.1 ± 2.7%	24.9 ± 1.3%	29.9 ± 1.0%	35.2 ± 0.9%	75.4 ± 1.8%
Depth 3	**Ours** gain, $K=7\uparrow$	19.0 ± 2.4%	26.1 ± 1.2%	32.1 ± 1.2%	38.6 ± 0.7%	76.6 ± 1.3%
Depth 3	Random ($K=7$)	+0.0730 ± 0.0859	+0.2617 ± 0.2130	+0.2750 ± 0.0765	+0.2773 ± 0.2665	+2.3928 ± 0.4840
Depth 3	Qwen3-8B	−0.0435 ± 0.1539 ✓	+0.1295 ± 0.1997 ✓	+0.0721 ± 0.0496 ✓	−0.0803 ± 0.1909 ✓	+2.0654 ± 0.2623 ✓
Depth 3	Llama-3.1-8B	+0.0883 ± 0.2356 ✗	+0.2766 ± 0.1085 ✗	+0.2083 ± 0.1798 ✓	+0.2084 ± 0.2366 ✓	+0.2603 ± 0.2330 ✓
Depth 3	DeepSeek-R1-8B	+0.0220 ± 0.1312 ✓	+0.2061 ± 0.0660 ✓	+0.1788 ± 0.1528 ✓	+0.3246 ± 0.1489 ✗	+1.6082 ± 0.4131 ✓
Depth 3	GPT-OSS-20B	+0.1633 ± 0.1815 ✗	+0.3438 ± 0.1193 ✗	+0.1786 ± 0.0759 ✓	+0.5089 ± 0.3596 ✗	+0.5647 ± 0.4252 ✓
Depth 3	**Ours** ($K=7$)	+0.0010 ± 0.0027 ✓	+0.0108 ± 0.0118 ✓	+0.0029 ± 0.0095 ✓	+0.0029 ± 0.0113 ✓	−0.0046 ± 0.0040 ✓
Depth 3	**Ours** ($K=1$)	**+0.0005 ± 0.0015** ✓	**+0.0092 ± 0.0131** ✓	**+0.0005 ± 0.0077** ✓	**−0.0014 ± 0.0097** ✓	**−0.0017 ± 0.0036** ✓

J.1 Clean Frozen-LLM Audit

Table 51: **Data-exposed frontier-model online VQE at depth 2.** The train/test split, $\tau = 0.995$, four-refresh $K=7$ protocol, and reporting conventions match Table 4.

Method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Energy fidelity ↓: ΔE vs. Full (Ha)					
Random ($K=7$)	+0.0930 ± 0.0297	+0.5668 ± 0.4463	+0.3054 ± 0.1633	+0.9327 ± 0.3273	+2.6969 ± 0.2816
GPT-5.4	+0.0032 ± 0.0064 ✓	+0.0090 ± 0.0164 ✓	+0.0011 ± 0.0056 ✓	**+0.0000 ± 0.0098** ✓	+0.0020 ± 0.0041 ✓
GPT-5.5	−0.0002 ± 0.0021 ✓	+0.0026 ± 0.0057 ✓	+0.0051 ± 0.0077 ✓	+0.0031 ± 0.0178 ✓	+0.0016 ± 0.0042 ✓
GPT-5.6 Sol	+0.0117 ± 0.0135 ✓	+0.0205 ± 0.0326 ✓	+0.0009 ± 0.0074 ✓	+0.0231 ± 0.0383 ✓	+0.0015 ± 0.0046 ✓
GPT-5.6 Terra	+0.0088 ± 0.0150 ✓	+0.0219 ± 0.0367 ✓	+0.0033 ± 0.0062 ✓	+0.0022 ± 0.0063 ✓	+0.0026 ± 0.0055 ✓
GPT-5.6 Luna	+0.0027 ± 0.0057 ✓	+0.0141 ± 0.0188 ✓	+0.0100 ± 0.0099 ✓	+0.0211 ± 0.0385 ✓	+0.0025 ± 0.0057 ✓
Claude Haiku 4.5	+0.0030 ± 0.0056 ✓	**−0.0001 ± 0.0013** ✓	+0.0030 ± 0.0070 ✓	+0.0146 ± 0.0514 ✓	**+0.0012 ± 0.0028** ✓
Claude Sonnet 5	+0.0003 ± 0.0003 ✓	−0.0001 ± 0.0009 ✓	+0.0016 ± 0.0062 ✓	−0.0039 ± 0.0115 ✓	+0.0043 ± 0.0092 ✓
Claude Opus 5	**+0.0000 ± 0.0005** ✓	+0.0001 ± 0.0003 ✓	+0.0054 ± 0.0073 ✓	−0.0031 ± 0.0186 ✓	+0.0037 ± 0.0082 ✓
Claude Fable 5	+0.0086 ± 0.0147 ✓	+0.0001 ± 0.0009 ✓	+0.0017 ± 0.0063 ✓	+0.0077 ± 0.0136 ✓	+0.0015 ± 0.0030 ✓
Ours ($K=7$)	−0.0007 ± 0.0015 ✓	+0.0076 ± 0.0176 ✓	**+0.0006 ± 0.0031** ✓	+0.0162 ± 0.0438 ✓	−0.0040 ± 0.0118 ✓
Measurement efficiency: shot gain (%; higher is better ↑)					
Random ($K=7$)	28.0 ± 2.2%	34.0 ± 1.4%	49.2 ± 1.4%	56.7 ± 0.8%	84.7 ± 0.6%
GPT-5.4	13.0 ± 0.3%	21.4 ± 1.3%	29.6 ± 1.0%	39.2 ± 1.2%	72.0 ± 1.8%
GPT-5.5	15.1 ± 0.8%	22.3 ± 1.6%	33.1 ± 0.3%	**60.6 ± 3.7%**	73.5 ± 0.5%
GPT-5.6 Sol	16.3 ± 1.3%	25.3 ± 1.8%	32.7 ± 2.1%	40.5 ± 0.8%	67.6 ± 0.8%
GPT-5.6 Terra	13.7 ± 1.8%	23.6 ± 2.6%	30.6 ± 0.6%	39.7 ± 1.2%	67.1 ± 1.0%
GPT-5.6 Luna	15.0 ± 0.9%	26.2 ± 1.3%	36.3 ± 1.5%	41.0 ± 1.2%	68.9 ± 1.3%
Claude Haiku 4.5	13.0 ± 1.6%	21.6 ± 1.0%	32.2 ± 1.2%	41.6 ± 1.9%	68.2 ± 1.0%
Claude Sonnet 5	11.6 ± 0.6%	21.3 ± 1.4%	29.5 ± 0.9%	57.6 ± 4.4%	77.4 ± 4.4%
Claude Opus 5	12.4 ± 1.0%	20.7 ± 1.0%	28.9 ± 0.7%	54.0 ± 3.6%	76.0 ± 2.1%
Claude Fable 5	14.6 ± 2.2%	21.1 ± 1.5%	29.4 ± 0.9%	29.0 ± 6.8%	66.9 ± 1.0%
Ours ($K=7$)	**27.3 ± 2.5%**	**32.5 ± 2.4%**	**47.9 ± 1.4%**	55.1 ± 0.5%	**83.7 ± 0.4%**

J.2 Frontier Models at Deeper Circuits

Table 52: **Data-exposed frontier-model online VQE at depth 3.** The train/test split, $\tau = 0.995$, four-refresh $K=7$ protocol, and reporting conventions match Table 4.

Method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
			Energy fidelity ↓: ΔE vs. Full (Ha)		
Random ($K=7$)	+0.0730 ± 0.0859	+0.2617 ± 0.2130	+0.2750 ± 0.0765	+0.2773 ± 0.2665	+2.3928 ± 0.4840
GPT-5.4	+0.0023 ± 0.0032 ✓	+0.0006 ± 0.0016 ✓	−0.0018 ± 0.0029 ✓	**+0.0019 ± 0.0031** ✓	+0.0018 ± 0.0076 ✓
GPT-5.5	+0.0010 ± 0.0036 ✓	−0.0003 ± 0.0027 ✓	+0.0007 ± 0.0036 ✓	+0.0134 ± 0.0121 ✓	−0.0012 ± 0.0122 ✓
GPT-5.6 Sol	+0.0014 ± 0.0025 ✓	**−0.0001 ± 0.0052** ✓	+0.0017 ± 0.0049 ✓	+0.0080 ± 0.0056 ✓	+0.0012 ± 0.0099 ✓
GPT-5.6 Terra	+0.0005 ± 0.0018 ✓	+0.0004 ± 0.0072 ✓	+0.0079 ± 0.0100 ✓	+0.0057 ± 0.0053 ✓	−0.0036 ± 0.0046 ✓
GPT-5.6 Luna	−0.0010 ± 0.0046 ✓	+0.0018 ± 0.0044 ✓	+0.0028 ± 0.0121 ✓	+0.0097 ± 0.0088 ✓	+0.0037 ± 0.0125 ✓
Claude Haiku 4.5	−0.0011 ± 0.0050 ✓	+0.0012 ± 0.0015 ✓	+0.0008 ± 0.0039 ✓	+0.0631 ± 0.1113 ✓	+0.0037 ± 0.0101 ✓
Claude Sonnet 5	−0.0021 ± 0.0020 ✓	+0.0001 ± 0.0014 ✓	+0.0016 ± 0.0011 ✓	+0.0131 ± 0.0133 ✓	−0.0026 ± 0.0122 ✓
Claude Opus 5	**−0.0003 ± 0.0009** ✓	−0.0005 ± 0.0032 ✓	+0.0004 ± 0.0035 ✓	+0.0095 ± 0.0099 ✓	+0.0023 ± 0.0109 ✓
Claude Fable 5	+0.0014 ± 0.0022 ✓	+0.0001 ± 0.0012 ✓	**+0.0001 ± 0.0009** ✓	+0.0019 ± 0.0045 ✓	**−0.0003 ± 0.0039** ✓
Ours ($K=7$)	+0.0010 ± 0.0027 ✓	+0.0108 ± 0.0118 ✓	+0.0029 ± 0.0095 ✓	+0.0029 ± 0.0113 ✓	−0.0046 ± 0.0040 ✓
			Measurement efficiency: shot gain (%; higher is better ↑)		
Random ($K=7$)	19.3 ± 1.6%	26.3 ± 1.0%	32.1 ± 1.4%	39.0 ± 0.9%	77.5 ± 0.9%
GPT-5.4	7.7 ± 1.5%	13.9 ± 1.3%	19.6 ± 0.8%	27.2 ± 2.3%	58.6 ± 1.9%
GPT-5.5	9.7 ± 0.4%	15.9 ± 0.8%	17.6 ± 0.3%	**39.8 ± 2.6%**	70.5 ± 2.7%
GPT-5.6 Sol	10.1 ± 0.6%	18.3 ± 0.7%	19.7 ± 1.4%	28.8 ± 0.7%	56.4 ± 0.5%
GPT-5.6 Terra	7.9 ± 0.1%	17.2 ± 0.7%	20.6 ± 1.6%	28.4 ± 0.8%	55.5 ± 1.0%
GPT-5.6 Luna	10.0 ± 1.1%	19.2 ± 0.7%	24.4 ± 0.9%	28.6 ± 0.5%	57.1 ± 2.5%
Claude Haiku 4.5	7.6 ± 1.3%	14.1 ± 1.1%	19.1 ± 0.9%	29.0 ± 0.9%	56.6 ± 0.8%
Claude Sonnet 5	7.5 ± 0.4%	12.5 ± 0.2%	18.0 ± 0.9%	28.1 ± 7.3%	70.6 ± 1.4%
Claude Opus 5	7.3 ± 0.8%	14.7 ± 1.0%	21.2 ± 0.9%	36.9 ± 3.5%	71.3 ± 1.2%
Claude Fable 5	9.8 ± 0.5%	13.5 ± 0.8%	17.2 ± 0.6%	19.1 ± 1.2%	55.8 ± 1.7%
Ours ($K=7$)	**19.0 ± 2.4%**	**26.1 ± 1.2%**	**32.1 ± 1.2%**	38.6 ± 0.7%	**76.6 ± 1.3%**

Table 53: **Online controller token traffic and API-equivalent cost.** The accounting covers the same 60 held-out VQE cells used in Tables 4–52: 28 optimizer steps per cell and four selector refreshes (240 total). Token categories come from run logs. OpenAI dollar entries are historical API-equivalent estimates retained from the original version; Claude entries are CLI-reported API totals. Dollar values describe historical API-equivalent accounting; token traffic is directly recorded in the experiment logs (Appendix N.1). Our method uses local checkpoint inference and no LLM tokens or API credits. Lower is better. Bold marks minima; underline marks the lowest nonzero entry in the total-token and cost columns.

Method	Token traffic (million)				API-equivalent cost (USD)	
	Non-cached input	Cached input	Output	Total	60-cell total	Per cell
GPT-5.4	12.71	3.64	0.78	17.13	44.43	0.74
GPT-5.5	12.58	4.42	0.45	17.45	78.67	1.31
GPT-5.6 Sol	16.50	1.46	0.57	18.52	100.35	1.67
GPT-5.6 Terra	13.46	3.62	0.61	17.70	43.76	0.73
GPT-5.6 Luna	14.74	1.74	0.52	<u>17.00</u>	<u>18.01</u>	<u>0.30</u>
Claude Haiku 4.5	24.58	1.08	6.75	32.40	71.29	1.19
Claude Sonnet 5	25.10	4.81	1.43	31.35	183.34	3.06
Claude Opus 5	24.90	2.75	0.55	28.20	159.01	2.65
Claude Fable 5	24.24	1.33	1.15	26.72	327.15	5.45
Ours	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**

Accounting note. OpenAI input totals include cached tokens, so the non-cached column subtracts the logged cached subset before pricing; Claude cache-read tokens are logged separately. The OpenAI reconstruction uses OpenAI API pricing, while Claude uses each CLI response's provider-reported `total_cost_usd` (see Anthropic pricing). These are normalized API-equivalent costs, not necessarily the subscription invoice. The one-time training cost of our method is excluded from this online-inference accounting; its 984,454-parameter checkpoint is 3.78 MiB and averages 7.29 ms per refresh on the local CPU.

Current takeaway

- Stronger frontier LLMs better integrate long structured contexts and exploit retrieved task supervision.
- No model consistently wins across molecules, depths, and fidelity–measurement trade-offs: current in-context policies are capable but not yet stable.
- The progression from clean prompts to physics context and data exposure shows substantial potential to identify task-relevant physical structure.

Future implication: Physics-aware agent

- Scale alone is insufficient; a physics-aware agent should combine retrieval, executable physics tools, memory, and verification.
- A promising hybrid uses our method as an efficient prior and invokes LLM reasoning only for OOD or low-confidence states.
- Crossed clean/data-exposed evaluations remain necessary before attributing the gain to frontier-model scale alone.

J.3 Observable Decision Traces

The main text retains the representative Llama trace. The complementary model traces below use the same fixed-state audit and expose only observable outputs, schema checks, selected-group evidence, and counterfactual feature tests.

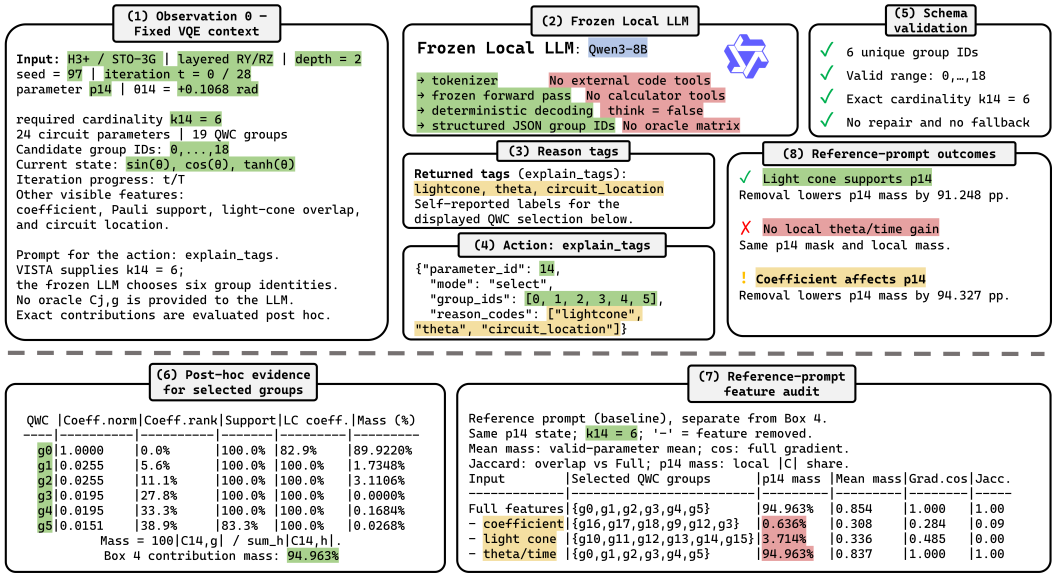

Figure 13: **Qwen3-8B observable decision trace.** On the fixed H_3^+, depth-2, seed-97 snapshot, Qwen selects $\{g_0, \ldots, g_5\}$ and captures 94.963% of the exact p_{14} contribution mass. Removing coefficient or light-cone features reduces that mass to 0.636% and 3.714%, respectively, whereas removing θ/time features leaves the reference mask unchanged. Box (4) uses the explanation prompt; Box (7) uses the separate reference prompt. Thus deletion measures reference-prompt selection quality. Raw reason tags are `lightcone`, `theta`, and `circuit_location`, as shown in Box (3). Box (8) reports reference-prompt utility. Exact contributions are audit-only (Appendix N.4).

How to read the trace. Boxes (1)–(5) show the complete model-visible interaction: a fixed VQE snapshot with 24 circuit parameters and 19 QWC groups is serialized into numerical descriptors, the frozen LLM returns reason tags and six group IDs, and a deterministic validator checks only schema, range, and cardinality. Boxes (6)–(8) are post-hoc scientific audits. They attach the exact contribution matrix only after the decision and delete feature families under the reference prompt. Prompt identity distinguishes the explanation action from its companion feature audit.

Observed decision. Qwen's executable mask $\{g_0, \ldots, g_5\}$ captures 94.963% of the local contribution mass without repair or fallback. The trace therefore separates two evaluation levels: executable output is assessed by the validator, and physical relevance is assessed by exact captured mass.

Reference-prompt feature utility

- Removing coefficients lowers captured mass from 94.963% to 0.636%, a 94.327 percentage-point loss.
- Removing light-cone features lowers it to 3.714%, a 91.248 percentage-point loss.

State sensitivity at this snapshot

- Removing θ/time leaves the selected p_{14} mask and its 94.963% local mass unchanged; mean mass across valid parameters changes from 0.854 to 0.837.
- The unchanged local mask coexists with strong coefficient and light-cone effects. The explanation action and reference-prompt sensitivities provide complementary observations.

Scientific interpretation. Qwen's reference selection responds strongly to physically meaningful coefficient and light-cone inputs, and its selected groups retain most of the exact row contribution. The aggregate online experiments extend this local view across molecules and optimization trajectories.

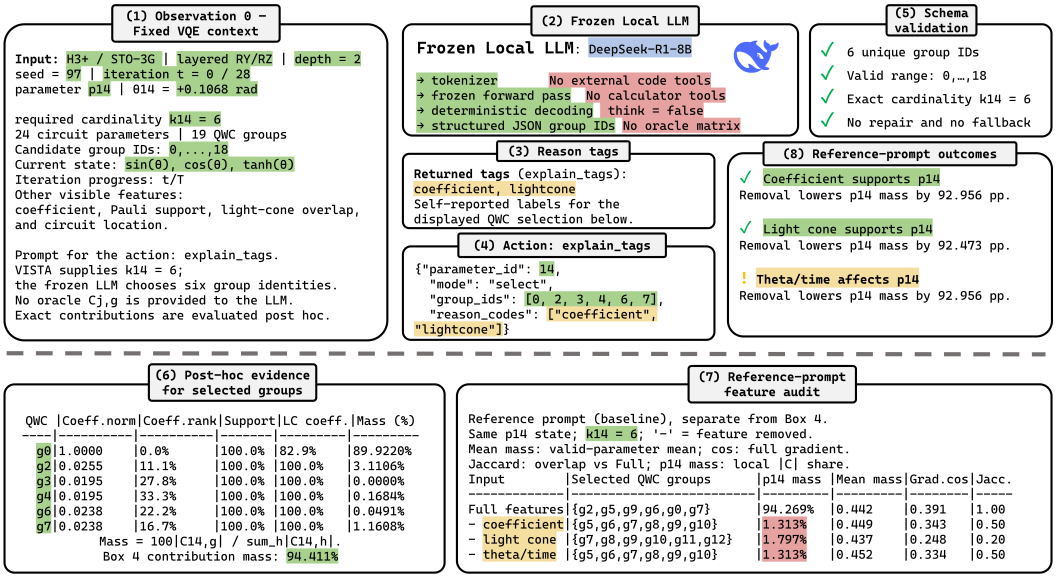

Figure 14: **DeepSeek-R1-8B observable decision trace.** DeepSeek selects $\{g_0, g_2, g_3, g_4, g_6, g_7\}$ and captures 94.411% of the exact p_{14} mass; its full-feature reference mask captures 94.269%. Removing coefficient or light-cone features reduces local mass to 1.313% and 1.797% under the separate reference prompt. Removing θ/time also reduces mass to 1.313%, identifying current-state sensitivity in the reference selection. Appendix N.4 records the explanation action and each reference-prompt intervention.

How to read the trace. The action panel contains only the IDs and observable reason tags emitted by the frozen model. Exact $C_{j,k}$ values in the evidence panel are never prompt features; they are attached afterward to score the action. The counterfactual panel tests reference-prompt feature utility alongside the separately logged explanation-producing action.

Observed decision. DeepSeek chooses $\{g_0, g_2, g_3, g_4, g_6, g_7\}$ and captures 94.411% of local mass, close to the 94.269% full-feature audit reference. Thus the model produces both a valid action and a high-quality local ranking on this snapshot.

Large reference-prompt quality effects

- Removing coefficients reduces local mass to 1.313% (a 92.956-point loss); removing light-cone features reduces it to 1.797% (a 92.473-point loss).
- Both input families support the high-quality local reference selection by retaining the dominant group g_0.

Broader feature sensitivity

- Removing θ/time under the reference prompt also reduces local mass to 1.313%. All three deletions omit the dominant group g_0.
- The reference ranking responds to current-state information as well as coefficients and light cones; the reported tags summarize the separate explanation action.

Scientific interpretation. DeepSeek produces a high-quality action on this cell, while the companion audit reveals sensitivity to several physical feature families. Together, the action score and intervention profile characterize the model's local selection behavior under the two documented prompt settings.

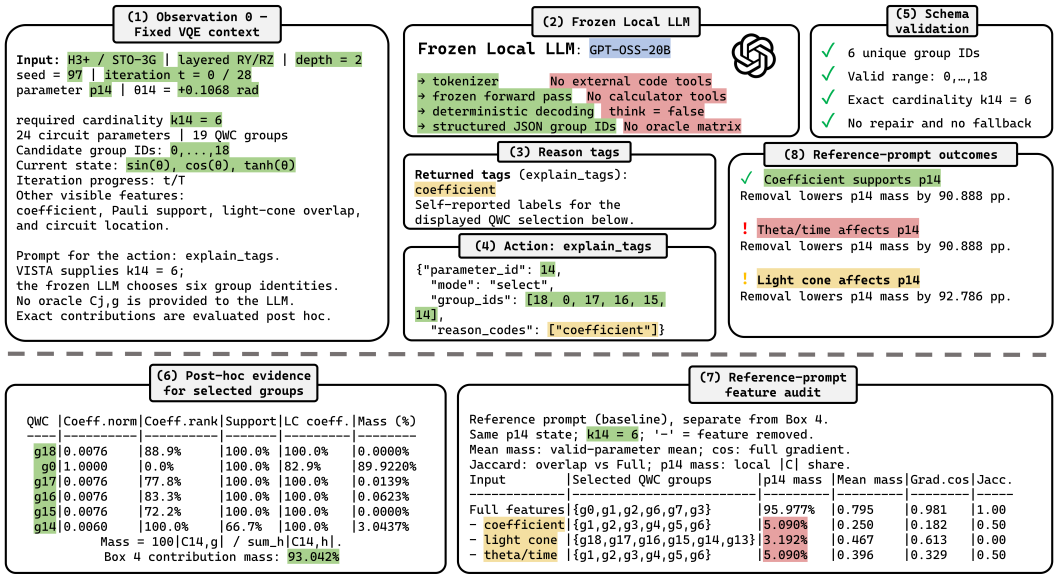

Figure 15: **GPT-OSS-20B observable decision trace.** GPT-OSS reports coefficient evidence and selects $\{g_{18}, g_0, g_{17}, g_{16}, g_{15}, g_{14}\}$, capturing 93.042% of the exact p_{14} mass; its reference mask captures 95.977%. Removing coefficient, light-cone, or θ/time features reduces local mass to 5.090%, 3.192%, and 5.090%, respectively, under the separate reference prompt. This ranking shows multi-feature quality sensitivity alongside the explanation action's coefficient tag (Appendix N.4).

How to read the trace. This example pairs a high-mass action with reference-prompt interventions. The frozen LLM generates selection JSON from serialized physical features; the external Python harness performs validation and all VQE simulation. The lower audit scores reference-prompt masks using exact contributions and feature-family deletion, with LLM tool execution disabled.

Observed decision. GPT-OSS selects $\{g_{18}, g_0, g_{17}, g_{16}, g_{15}, g_{14}\}$ and captures 93.042% of local mass. This is a strong mask, alongside the 95.977% full-feature reference. Both actions concentrate measurements on groups carrying most of the exact parameter-row contribution.

Reference-prompt feature utility

- Removing coefficients lowers local captured mass to 5.090%, a 90.888 percentage-point loss.
- Coefficients improve the reference-prompt selection. The separately prompted explanation action also reports a coefficient tag.

Multiple sensitive input families

- Removing light-cone or θ/time features lowers mass to 3.192% and 5.090%, respectively.
- The reference ranking draws useful information from several physical families, complementing the coefficient tag returned under the explanation prompt.

Scientific interpretation. The deletion pattern is consistent with joint integration of several serialized physical inputs. The external harness evaluates this behavior through captured mass and feature sensitivity, with model-generated output restricted to selection JSON. The successful action and its companion audit provide two concrete views of GPT-OSS's physical selection capability.

Table 54: **Prompt, output, and context ablations with token cost.** Values are macro-averaged across Qwen3-8B, Llama-3.1-8B, DeepSeek-R1-8B, and GPT-OSS-20B on the 12 fixed audit snapshots per model. Token and call costs include retries, context splits, and invalid outputs, and are divided by all 48 attempted model–snapshot cases. Performance metrics are averaged over protocol-successful cases, whose identities can change across conditions; this is not an all-case matched accuracy comparison. Each ablation is compared with the first-row 32K reference; that reference is repeated in the context block for a direct 8K/16K/32K comparison. Within each numeric column, **bold** marks the best reported value and underlining marks the second-best distinct value, following the indicated arrow; tied values share the same formatting, and the repeated reference is counted once. These column-wise marks do not imply one overall-best protocol. The full-ranking row (†) succeeds in only 22/48 cases and is survivor-biased.

Ablation axis	Setting	Captured mass ↑	Grad. cosine ↑	Success ↑	Prompt kTok ↓	Output kTok ↓	Calls ↓
Reference	Domain + compressed, 32K	0.554	0.628	97.9%	160.2	12.0	10.1
Prompt	Minimal + compressed, 32K	0.549	0.655	**100.0%**	140.6	8.7	9.0
	Domain + 2-shot, 32K	0.510	0.509	97.9%	189.1	13.4	10.8
Output	Domain + direct IDs, 32K	0.594	0.633	95.8%	169.8	13.2	11.5
	Domain + full ranking, 32K†	**0.677**	**0.789**	45.8%	114.6	16.5	12.5
Context	Domain + compressed, 8K	0.562	0.592	60.4%	**56.7**	**5.9**	**8.0**
	Domain + compressed, 16K	0.564	0.583	81.2%	113.5	10.2	9.5
	Domain + compressed, 32K	0.554	0.628	97.9%	160.2	12.0	10.1

J.4 Prompt and Context Cost

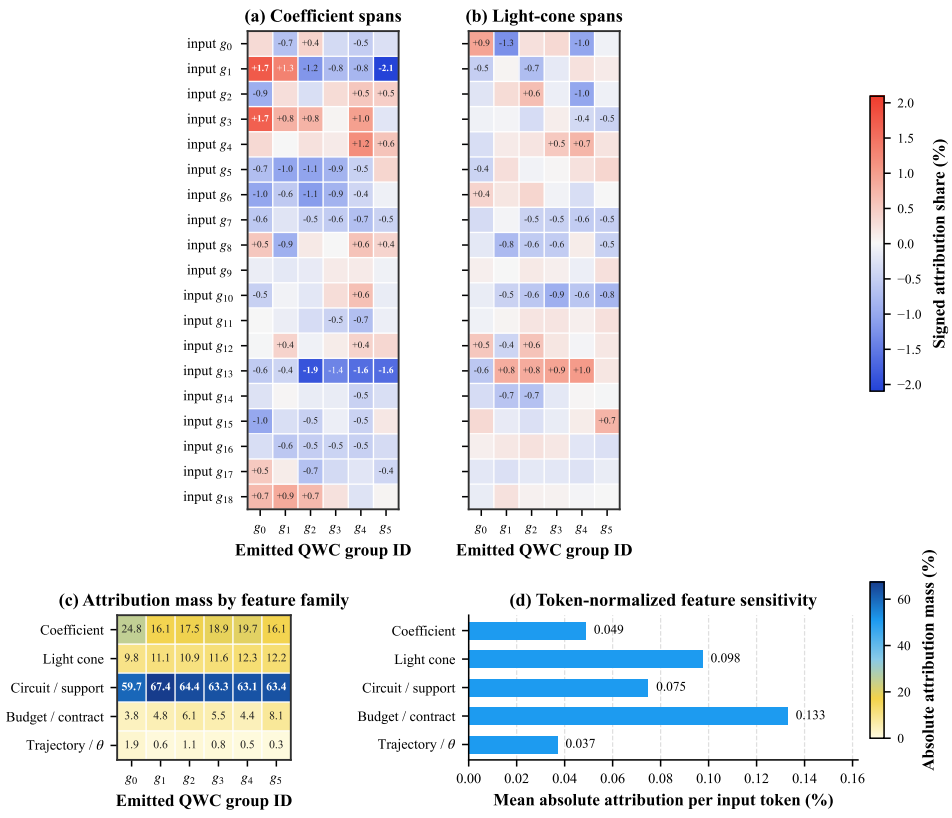

Figure 16: **Token-to-QWC-group attribution for one representative frozen-Qwen decision.** The audited state is H_3^+ with STO-3G, depth 2, seed 97, iteration 0, and parameter p_{14}; our method supplies $k_{14} = 6$, and Qwen emits $[g_0, g_1, g_2, g_3, g_4, g_5]$. Panels (a,b) report the signed share of integrated-gradient attribution from each input group's coefficient and parameter–group light-cone spans to each emitted ID: red raises and blue lowers the grammar-conditioned target score relative to the numeric-zero baseline. Panel (c) aggregates absolute attribution mass by feature family, while panel (d) divides that mass by the number of input tokens in each family. The observable action comes from the audit-panel Ollama `qwen3:8b` run; gradients are computed with the official frozen `Qwen/Qwen3-8B` checkpoint (Yang et al., 2025) in NF4 4-bit mode because the Ollama/GGUF runtime does not expose input gradients. This matched white-box surrogate is not bitwise identical to the generation runtime. The 32-point completeness audit has maximum residual 0.122; for the informative first decision g_0, this is 1.61% of its score change, but later columns have large relative errors. The plot attributes the teacher-forced output score; Appendix N.5 reports the per-column numerical audit.

Key finding. The saved estimates emphasize circuit/support, coefficient, and light-cone spans on this snapshot. The first decision has a 1.61% relative completeness residual; later continuation scores have larger relative residuals. The displayed profile describes this fixed-prefix attribution calculation.

J.5 Token- and Feature-Level Attribution

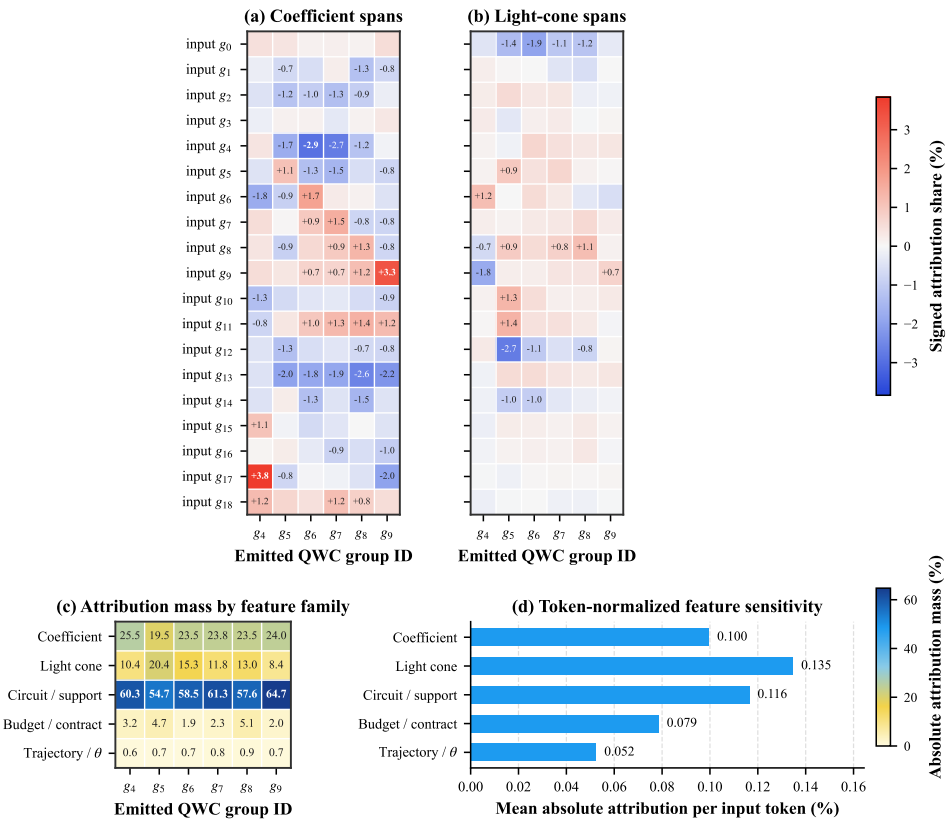

Figure 17: **Token-to-QWC-group attribution for Llama-3.1-8B.** The state and panel definitions match Figure 16, while the accepted reference-prompt action is $[g_4, g_5, g_6, g_7, g_8, g_9]$. The observable action comes from Ollama `llama3.1:8b`; gradients use a public NF4 conversion of the frozen Meta-Llama-3.1-8B-Instruct checkpoint (Grattafiori et al., 2024), which is not bitwise identical to the GGUF generation runtime. The 32-point maximum completeness residual is 1.149 log-probability units. Appendix N.5 reports absolute and relative numerical residuals alongside the estimates.

Key finding. The saved Llama map reports a 0.73% trajectory share with the numerical residuals tabulated in Appendix N.5. Its first ID g_4 has only 1.06% conditional probability (rank 15), whereas g_5–g_9 become rank-one continuations, exposing a strong autoregressive list-completion pattern in the white-box runtime.

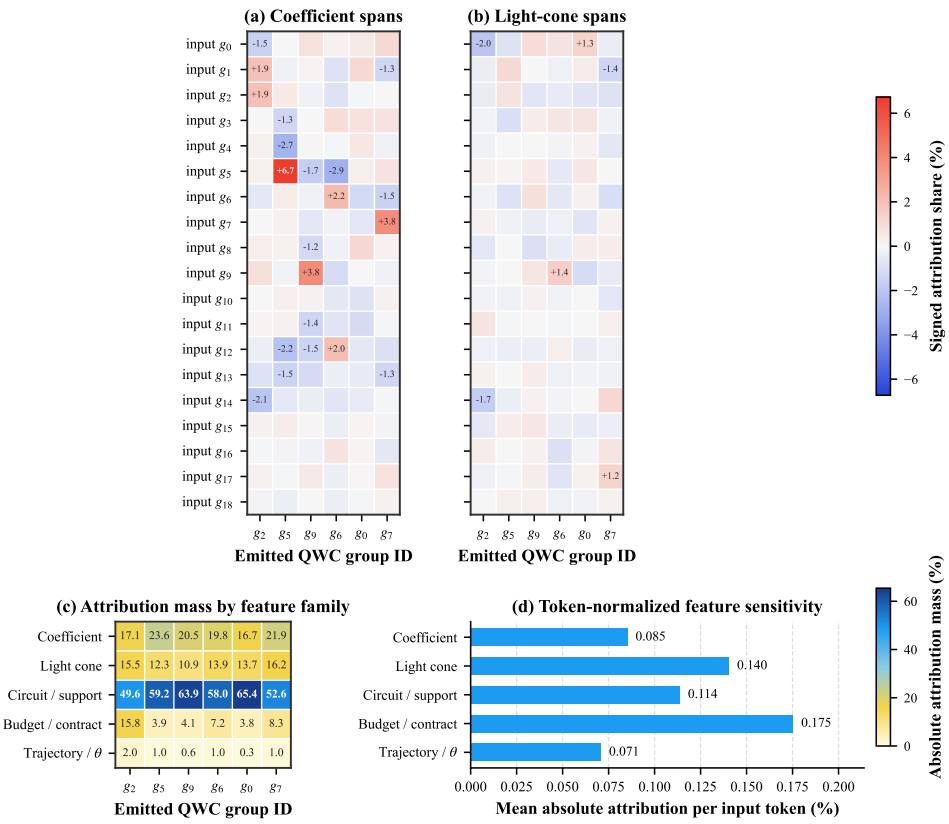

Figure 18: **Token-to-QWC-group attribution for DeepSeek-R1-Distill-Llama-8B.** The accepted reference-prompt action is $[g_2, g_5, g_9, g_6, g_0, g_7]$ under the same H_3^+ snapshot and attribution protocol. The action comes from Ollama `deepseek-r1:8b`; white-box gradients use the official frozen DeepSeek-R1-Distill-Llama-8B checkpoint (Guo et al., 2025) with runtime NF4 quantization. This surrogate is not bitwise identical to the GGUF runtime. The 32-point maximum completeness residual is 0.437 log-probability units; absolute and relative numerical residuals are reported in Appendix N.5.

Key finding. The saved DeepSeek map emphasizes static structure, coefficients, and light-cone spans. The accepted IDs receive only 0.14–30.85% conditional probability and ranks 2–15 in the white-box runtime. The panel records input sensitivity of these fixed-prefix scores, with runtime settings and integration residuals specified by the accompanying audit.

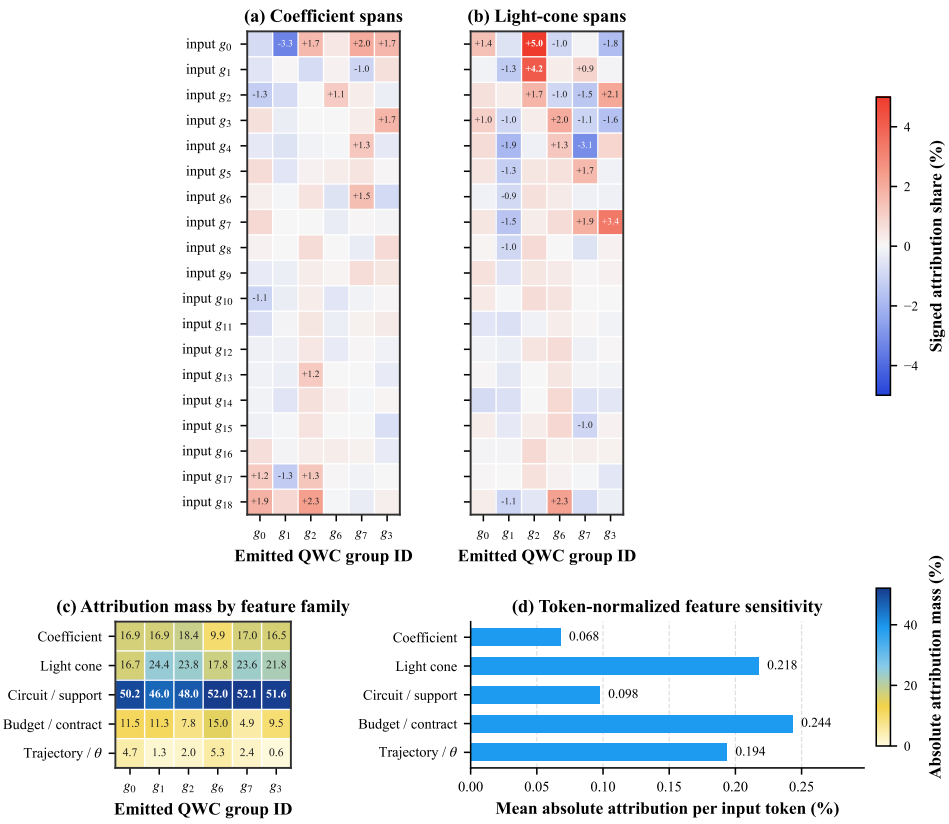

Figure 19: **Token-to-QWC-group attribution for GPT-OSS-20B.** The accepted reference-prompt action is $[g_0, g_1, g_2, g_6, g_7, g_3]$ under the same snapshot and 32-point path. The action comes from Ollama `gpt-oss:20b`; gradients use the official frozen OpenAI GPT-OSS-20B native-MXFP4 checkpoint (Agarwal et al., 2025). Flex Attention and one forward graph per output bound memory without changing the scalar scores or quadrature nodes. The runtime is not bitwise identical to Ollama, and the measured maximum completeness residual is 1.414 log-probability units; the full numerical audit appears in Appendix N.5.

Key finding. GPT-OSS's saved light-cone estimate is 21.36%. Four emitted IDs are high-confidence while g_6 and g_3 remain below 0.4%, indicating a heterogeneous fixed-prefix score profile in the white-box runtime.

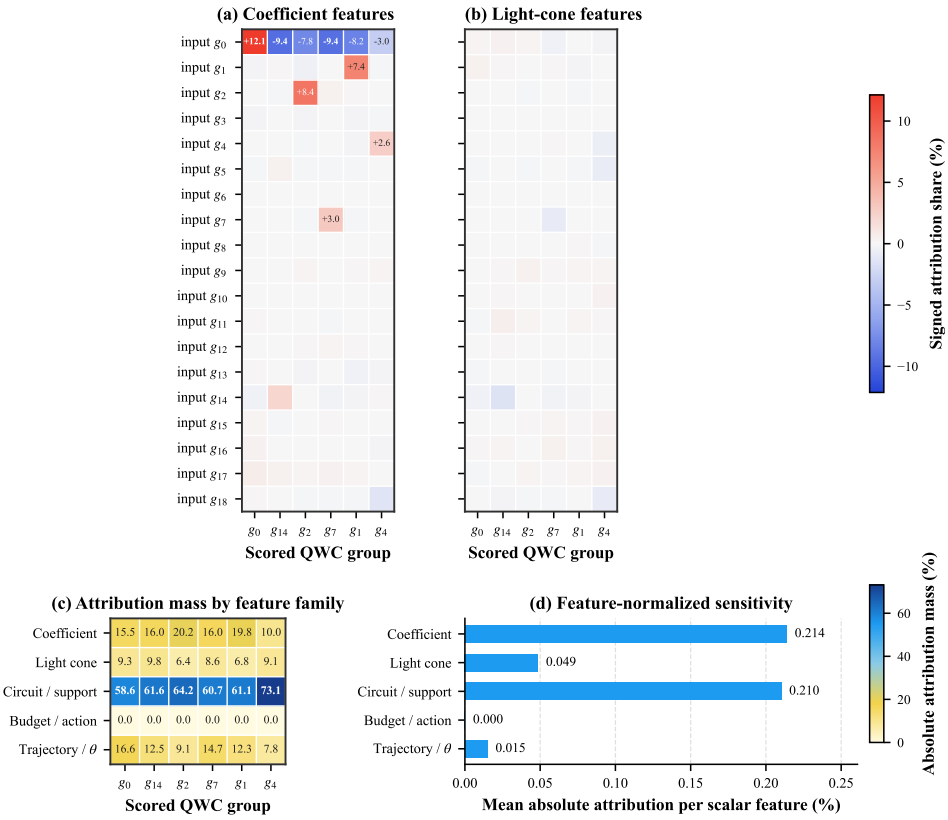

Figure 20: **Model-native feature-to-QWC attribution for the trained controller.** We freeze the saved Transformer checkpoint and use the same H_3^+/STO-3G, depth-2, seed-97, iteration-0, parameter-p_{14} snapshot as in Figures 16–19. The six displayed columns are our method's highest-scoring groups $(g_0, g_{14}, g_2, g_7, g_1, g_4)$, fixing the LLM comparison cardinality; the native rounded budget head predicts five groups. Panels (a) and (b) trace coefficient and light-cone inputs from every QWC token to each scored group, (c) aggregates absolute attribution by feature family, and (d) normalizes that mass by the number of scalar features in each family. Attributions use 1,024-point Gauss–Legendre integrated gradients from the training-mean baseline in standardized feature space and include all parameter tokens, group tokens, and parameter–group pair features. The maximum completeness residual is 0.0126 log-probability units, with maximum relative residual 0.46%. This panel measures native feature-score dependence; the LLM panels measure token-score dependence under their respective baselines.

Key finding. Our method combines strong circuit/support dependence (63.24%) with explicit use of trajectory features (12.16%). Its top-six ranking contains all five groups in the exact $\tau = 0.995$ top-mass mask. The native-checkpoint attribution connects current optimization state to the scores used for parameter-specific measurement selection.

Table 55: **Does the selected prompt improve clean frozen-LLM selection online?** GPT-OSS (clean) is the clean exact select/exclude baseline; GPT-OSS (best) uses the selected domain prompt, direct-ID output, 32K context, and all physical features. Both remain frozen, receive the same count k_j requested by our method, and refresh group identities every $K=7$ steps. Gain and energy entries are mean $\pm$ standard deviation over seeds 97/101/107/109, with ΔE in Ha relative to Full-QWC; gradients use exact expectations and shot costs are nominal (Appendix N.1). An energy cell for the best prompt is green when its absolute mean gap is smaller than that of the clean baseline and red otherwise. The complete Random and our method controls are reported in Appendix Table 57.

Setting	Protocol	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 1	GPT clean gain, $K=7\uparrow$	$39.4 \pm 1.8\%$	$62.6 \pm 1.3\%$	$67.0 \pm 1.5\%$	$79.6 \pm 1.8\%$	$89.9 \pm 0.8\%$
Depth 1	GPT best gain, $K=7\uparrow$	$42.5 \pm 4.6\%$	$61.7 \pm 3.4\%$	$66.7 \pm 1.5\%$	$79.3 \pm 1.8\%$	$90.5 \pm 0.4\%$
Depth 1	GPT-OSS (clean), ΔE	$+0.3930 \pm 0.1936$	$+0.8238 \pm 0.2303$	$+0.7300 \pm 0.4383$	$+0.5671 \pm 0.3164$	$+1.5503 \pm 0.7644$
Depth 1	GPT-OSS (best), ΔE	$+0.3392 \pm 0.1856$	$+0.9079 \pm 0.4907$	$+0.3882 \pm 0.2525$	$+0.6337 \pm 0.2763$	$+1.9964 \pm 0.9332$
Depth 2	GPT clean gain, $K=7\uparrow$	$29.5 \pm 3.5\%$	$34.3 \pm 1.7\%$	$49.8 \pm 1.9\%$	$55.5 \pm 1.6\%$	$84.1 \pm 0.5\%$
Depth 2	GPT best gain, $K=7\uparrow$	$29.7 \pm 2.2\%$	$33.0 \pm 2.3\%$	$49.8 \pm 0.8\%$	$55.4 \pm 0.4\%$	$84.3 \pm 0.5\%$
Depth 2	GPT-OSS (clean), ΔE	$+0.2148 \pm 0.1220$	$+0.6807 \pm 0.7081$	$+0.4691 \pm 0.2024$	$+0.5616 \pm 0.2675$	$+1.2894 \pm 0.3290$
Depth 2	GPT-OSS (best), ΔE	-0.0510 ± 0.0402	$+0.1302 \pm 0.0822$	$+0.4776 \pm 0.3130$	$+0.4682 \pm 0.1297$	$+1.5294 \pm 0.3237$
Depth 3	GPT clean gain, $K=7\uparrow$	$19.8 \pm 1.9\%$	$25.1 \pm 1.8\%$	$32.1 \pm 1.4\%$	$39.0 \pm 0.5\%$	$76.8 \pm 1.5\%$
Depth 3	GPT best gain, $K=7\uparrow$	$19.5 \pm 1.6\%$	$26.3 \pm 1.2\%$	$32.1 \pm 1.0\%$	$38.7 \pm 1.1\%$	$76.6 \pm 1.3\%$
Depth 3	GPT-OSS (clean), ΔE	$+0.1633 \pm 0.1815$	$+0.3438 \pm 0.1193$	$+0.1786 \pm 0.0759$	$+0.5089 \pm 0.3596$	$+0.5647 \pm 0.4252$
Depth 3	GPT-OSS (best), ΔE	$+0.0478 \pm 0.0637$	$+0.0162 \pm 0.1316$	$+0.1222 \pm 0.1216$	$+0.0449 \pm 0.2110$	$+0.5331 \pm 0.3351$

J.6 Best-Prompt Online VQE and Explanation Faithfulness

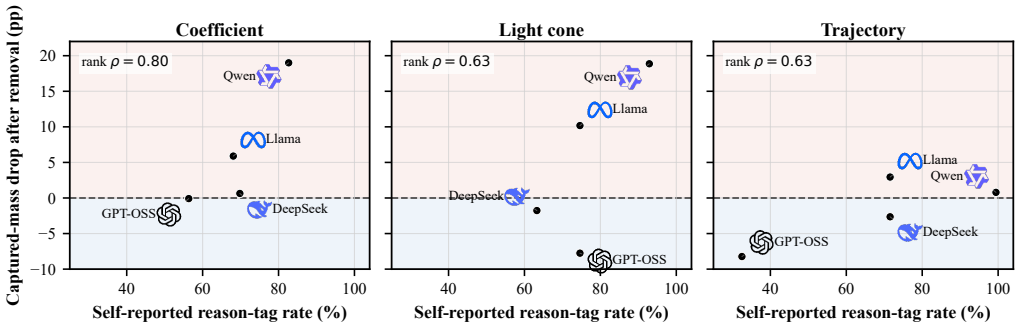

Figure 21: **Self-reported reasons versus measured perturbation sensitivity.** We audit three input families: coefficient denotes static Hamiltonian-coefficient summaries for each QWC group; light cone denotes parameter–group support and coefficient overlap; and trajectory denotes only the current optimization-state proxy, not a stored history. Specifically, for parameter j it contains $[\sin\theta_j, \cos\theta_j, \tanh\theta_j, t/T]$, together with prompt-level iteration t and total iterations T. It does not include previous energies, gradients, Adam moments, measurement outcomes, or earlier masks. For each frozen LLM and feature family, the horizontal axis is the fraction of explained rows that name the corresponding reason code (the trajectory rate combines θ and iteration tags). The vertical axis is the paired decrease in exact captured contribution mass when that entire feature family is removed from the same Full-QWC snapshot, without retraining or changing the requested cardinality. Positive values indicate beneficial input utility under the reference prompt; negative values indicate harmful utility. Near-zero effects may reflect redundancy. Reason tags come from a separate explanation prompt, giving a cross-prompt association between reported tags and measured feature utility. Black points are measured model–feature pairs and adjacent labels carry the model-family logo. The rank correlation is descriptive over only four models, with prompt provenance in Appendix N.4.

Table 56: **What drives frozen-LLM QWC selection?** The audit evaluates four frozen local LLMs on the same 12 Full-QWC snapshots (H3+, H4 chain, and LiH; depths 2 and 3; seed 97; iterations 0 and 14), with no fine-tuning and identical per-parameter cardinalities. The reference uses a domain prompt, compressed output, 32K context, and all non-oracle coefficient, light-cone, parameter, and trajectory features. Δ captured mass is the paired, model-macro-averaged change from this reference in percentage points; positive is better. Protocol success is the fraction of the 48 model–snapshot cases yielding a valid exact-cardinality mask. The apparent ranking-output gain is conditional on only 22 successful cases and should not be interpreted as a reliable improvement.

Factor	Intervention	Δ mass (pp) ↑	Protocol success ↑	Main observation
Prompt	Minimal instruction	-1.0	100.0%	Shorter wording improves validity but not relevance quality.
Prompt	Two demonstrations	-4.8	97.9%	Few-shot prompting does not provide a universal gain.
Output	Direct selected IDs	$\mathbf{+3.5}$	95.8%	The strongest reliable output contract on average.
Output	Full ranking	$+7.9^{\dagger}$	45.8%	Long structured outputs are brittle and strongly survivor-biased.
Context	8K tokens	-1.2	60.4%	Frequent context splitting and incomplete structured outputs.
Context	16K tokens	-2.4	81.2%	Higher reliability than 8K, without a quality gain on valid cases.
Context	32K reference	0.0	**97.9%**	Long context primarily prevents protocol failure.
Physics	Remove coefficients	$\mathbf{-6.4}$	97.9%	Coefficient information is useful overall, especially to Qwen and Llama.
Physics	Remove light-cone features	$\mathbf{-4.9}$	93.8%	Light-cone information is useful overall but strongly model-dependent.
Physics	Remove θ/iteration features	$+1.8$	97.9%	No stable aggregate benefit from current-state fields.
Physics	Shuffle coefficients	-1.5	91.7%	The largest shuffle effect is in Qwen rather than universal.
Physics	Shuffle light-cone features	$+2.3$	93.8%	Opposite model-level effects cancel in the aggregate.

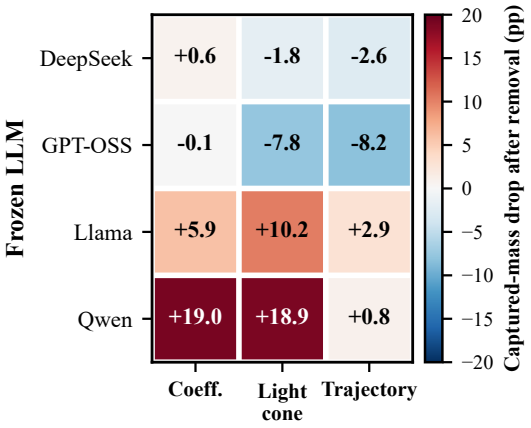

Figure 22: **Frozen LLMs use physical feature families differently.** Each cell is the paired captured-mass drop, in percentage points, after removing one feature family from the 32K full-feature reference. Positive red values indicate useful dependence; negative blue values indicate that removal improves selection. The aggregate benefit is model dependent, and per-snapshot effects may reverse sign.

J.7 Complete Mechanism-Audit Summary

Table 56 preserves the aggregate intervention summary behind the main-text attribution and cost views.

J.8 Complete Best-Prompt Online-VQE Controls

Table 57 retains the Random baseline and the refresh-matched and per-step our method controls omitted from the compact clean-versus-best comparison in the main text.

Table 57: **Complete best-prompt closed-loop online-VQE comparison.** This appendix table preserves the full matched-control version of Table 55. GPT-OSS (clean) is the clean exact select/exclude baseline, while GPT-OSS (best) uses the audit-selected domain prompt, direct-ID output, 32K context, and all physical features. Both models remain frozen, receive only the count k_j requested by the proposed method, and refresh group identities every $K=7$ steps. Random uses the same cardinality and refresh cadence; our method is reported at both $K=7$ and per-step $K=1$. Gain rows report realized mean ± standard deviation over seeds 97/101/107/109; all safety fallbacks pay Full-QWC shots. Energy entries are mean ± standard deviation relative to Full-QWC. Green/red satisfies/violates the one-sided 1.6 mHa no-worse criterion; the black check/cross compares absolute mean gap with Random. Bold/underline mark the smallest/second-smallest absolute mean gap among the five methods in each molecule–depth cell.

Setting	Selection method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 1	GPT clean gain, $K=7\uparrow$	39.4 ± 1.8%	62.6 ± 1.3%	67.0 ± 1.5%	79.6 ± 1.8%	89.9 ± 0.8%
Depth 1	GPT best gain, $K=7\uparrow$	42.5 ± 4.6%	61.7 ± 3.4%	66.7 ± 1.5%	79.3 ± 1.8%	90.5 ± 0.4%
Depth 1	**Ours** gain, $K=7\uparrow$	41.2 ± 5.3%	60.7 ± 3.1%	65.1 ± 2.6%	77.3 ± 0.7%	88.2 ± 0.4%
Depth 1	**Ours** gain, $K=1\uparrow$	40.4 ± 6.3%	58.9 ± 3.5%	63.0 ± 2.8%	74.7 ± 1.1%	86.9 ± 0.5%
Depth 1	Random ($K=7$)	+0.1847 ± 0.0868	+1.2191 ± 0.2759	+0.7247 ± 0.2856	+0.9736 ± 0.3101	+2.9030 ± 0.6058
Depth 1	GPT-OSS (clean)	+0.3930 ± 0.1936 ✗	+0.8238 ± 0.2303 ✓	+0.7300 ± 0.4383 ✗	+0.5671 ± 0.3164 ✓	+1.5503 ± 0.7644 ✓
Depth 1	GPT-OSS (best)	+0.3392 ± 0.1856 ✗	+0.9079 ± 0.4907 ✓	+0.3882 ± 0.2525 ✓	+0.6337 ± 0.2763 ✓	+1.9964 ± 0.9332 ✓
Depth 1	**Ours** ($K=7$)	+0.0013 ± 0.0003 ✓	+0.0035 ± 0.0015 ✓	**+0.0004 ± 0.0006** ✓	**+0.0001 ± 0.0060** ✓	+0.0030 ± 0.0016 ✓
Depth 1	**Ours** ($K=1$)	**+0.0011 ± 0.0008** ✓	**+0.0015 ± 0.0012** ✓	+0.0005 ± 0.0006 ✓	−0.0006 ± 0.0024 ✓	**+0.0023 ± 0.0016** ✓
Depth 2	GPT clean gain, $K=7\uparrow$	29.5 ± 3.5%	34.3 ± 1.7%	49.8 ± 1.9%	55.5 ± 1.6%	84.1 ± 0.5%
Depth 2	GPT best gain, $K=7\uparrow$	29.7 ± 2.2%	33.0 ± 2.3%	49.8 ± 0.8%	55.4 ± 0.4%	84.3 ± 0.5%
Depth 2	**Ours** gain, $K=7\uparrow$	27.3 ± 2.5%	32.5 ± 2.4%	47.9 ± 1.4%	55.1 ± 0.5%	83.7 ± 0.4%
Depth 2	**Ours** gain, $K=1\uparrow$	25.6 ± 2.9%	31.6 ± 2.6%	44.9 ± 1.3%	51.8 ± 0.5%	82.6 ± 0.4%
Depth 2	Random ($K=7$)	+0.0930 ± 0.0297	+0.5668 ± 0.4463	+0.3054 ± 0.1633	+0.9327 ± 0.3273	+2.6969 ± 0.2816
Depth 2	GPT-OSS (clean)	+0.2148 ± 0.1220 ✗	+0.6807 ± 0.7081 ✗	+0.4691 ± 0.2024 ✗	+0.5616 ± 0.2675 ✓	+1.2894 ± 0.3290 ✓
Depth 2	GPT-OSS (best)	−0.0510 ± 0.0402 ✓	+0.1302 ± 0.0822 ✓	+0.4776 ± 0.3130 ✗	+0.4682 ± 0.1297 ✓	+1.5294 ± 0.3237 ✓
Depth 2	**Ours** ($K=7$)	−0.0007 ± 0.0015 ✓	+0.0076 ± 0.0176 ✓	**+0.0006 ± 0.0031** ✓	+0.0162 ± 0.0438 ✓	**−0.0040 ± 0.0118** ✓
Depth 2	**Ours** ($K=1$)	**+0.0003 ± 0.0007** ✓	**−0.0011 ± 0.0026** ✓	−0.0015 ± 0.0041 ✓	**+0.0158 ± 0.0443** ✓	−0.0049 ± 0.0097 ✓
Depth 3	GPT clean gain, $K=7\uparrow$	19.8 ± 1.9%	25.1 ± 1.8%	32.1 ± 1.4%	39.0 ± 0.5%	76.8 ± 1.5%
Depth 3	GPT best gain, $K=7\uparrow$	19.5 ± 1.6%	26.3 ± 1.2%	32.1 ± 1.0%	38.7 ± 1.1%	76.6 ± 1.3%
Depth 3	**Ours** gain, $K=7\uparrow$	19.0 ± 2.4%	26.1 ± 1.2%	32.1 ± 1.2%	38.6 ± 0.7%	76.6 ± 1.3%
Depth 3	**Ours** gain, $K=1\uparrow$	18.1 ± 2.7%	24.9 ± 1.3%	29.9 ± 1.0%	35.2 ± 0.9%	75.4 ± 1.8%
Depth 3	Random ($K=7$)	+0.0730 ± 0.0859	+0.2617 ± 0.2130	+0.2750 ± 0.0765	+0.2773 ± 0.2665	+2.3928 ± 0.4840
Depth 3	GPT-OSS (clean)	+0.1633 ± 0.1815 ✗	+0.3438 ± 0.1193 ✗	+0.1786 ± 0.0759 ✓	+0.5089 ± 0.3596 ✗	+0.5647 ± 0.4252 ✓
Depth 3	GPT-OSS (best)	+0.0478 ± 0.0637 ✓	+0.0162 ± 0.1316 ✓	+0.1222 ± 0.1216 ✓	+0.0449 ± 0.2110 ✓	+0.5331 ± 0.3351 ✓
Depth 3	**Ours** ($K=7$)	+0.0010 ± 0.0027 ✓	+0.0108 ± 0.0118 ✓	+0.0029 ± 0.0095 ✓	+0.0029 ± 0.0113 ✓	−0.0046 ± 0.0040 ✓
Depth 3	**Ours** ($K=1$)	**+0.0005 ± 0.0015** ✓	**+0.0092 ± 0.0131** ✓	**+0.0005 ± 0.0077** ✓	**−0.0014 ± 0.0097** ✓	**−0.0017 ± 0.0036** ✓

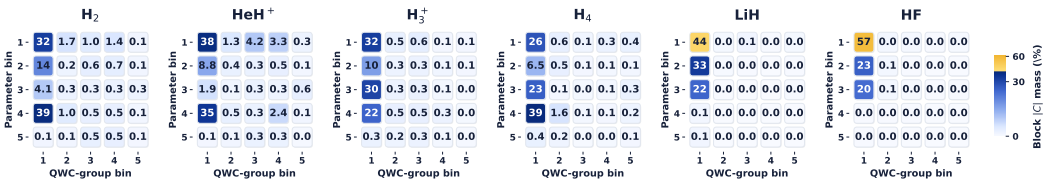

Figure 23: **The same contribution matrices under IBM calibration-derived noise.** Molecules, circuits, initialization, parameter shifts, block aggregation, and color normalization are identical to Figure 1. We extract median gate and readout errors, gate durations, and coherence metadata from the fixed Qiskit `FakeGuadalupeV2` system snapshot (Javadi-Abhari et al., 2024; IBM Quantum, 2026a). Following standard backend-derived noise modeling (IBM Quantum, 2026b), our scalable Pauli-channel surrogate combines gate depolarization, T_2-limited decay, and readout attenuation. This is a calibrated noise simulation, not execution on a physical IBM QPU.

K Calibration-Derived Noise Diagnostics

Across 120 surrogate snapshots, clean and noisy contribution grids have 50.6% and 51.5% mean sparsity, respectively. Their masks have 0.970 mean Jaccard overlap, their relevance rankings have 0.996 mean Spearman correlation, and a clean top-mass mask retains 99.7% of noisy contribution mass. The smaller gate-level Aer check is deliberately more conservative: finite-shot transpiled circuits reduce mask overlap and rank correlation, but 15–26% of parameter–group cells remain unselected at $\tau = 0.995$. Neither experiment is a physical-QPU run.

Table 58: **Noisy online VQE comparison at $\tau = 0.995$.** This repeats the Table 1 protocol under a finite-shot Pauli-channel surrogate derived from the fixed IBM FakeGuadalupeV2 calibration snapshot (Javadi-Abhari et al., 2024; IBM Quantum, 2026a;b). Final energies are evaluated noiselessly and paired against noisy Full-QWC training; † rows retain the noiseless cell-wise TITAN thresholds. The same noise study additionally tests contribution-mask stability (Figure 23 and Table 59), a gate-level Aer implementation (Table 60), and Hamiltonian-family transfer (Table 61).

Setting	Metric	Method	H2 (4q)	HeH+ (4q)	H3+ (6q)	H4 chain (8q)	LiH (12q)
Depth 1	ΔE vs Full ↓	**Ours**	−0.0076	−0.0219	−0.0252	+0.0001	−0.0304
Depth 1	ΔE vs Full ↓	Random	+0.2170	+1.2823	+0.9091	+1.3470	+3.3622
Depth 1	ΔE vs Full ↓	Global coeff.	−0.0050	+0.0155	−0.0022	−0.0020	+0.0041
Depth 1	ΔE vs Full ↓	TITAN (Peng et al.)	+0.0150	+0.0723	+0.0413	+0.0116	−0.1104
Depth 1	ΔE vs Full ↓	TITAN+**Ours**	+0.0138	+0.0746	+0.0422	+0.0138	−0.1065
Depth 1	ΔE vs Full ↓	TITAN†	+0.0053	+0.0985	+0.0293	−0.0851	−0.0705
Depth 1	ΔE vs Full ↓	TITAN†+**Ours**	+0.0011	+0.0988	+0.0291	−0.0864	−0.0693
Depth 1	Shot gain ↑	**Ours**	43.2%	59.9%	64.0%	75.9%	88.1%
Depth 1	Shot gain ↑	Random	42.6%	60.5%	67.9%	79.1%	90.3%
Depth 1	Shot gain ↑	Global coeff.	41.5%	61.1%	64.8%	76.1%	88.3%
Depth 1	Shot gain ↑	TITAN (Peng et al.)	65.0%	56.2%	58.6%	63.1%	73.7%
Depth 1	Shot gain ↑	TITAN+**Ours**	**81.7%**	90.5%	93.6%	96.1%	99.7%
Depth 1	Shot gain ↑	TITAN†	43.6%	59.4%	64.3%	71.6%	82.3%
Depth 1	Shot gain ↑	TITAN†+**Ours**	65.9%	**91.7%**	**95.5%**	**98.0%**	**99.9%**
Depth 1	No-worse vs Full	**Ours**	✓	✓	✓	✓	✓
Depth 1	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 1	No-worse vs Full	Global coeff.	✓	✗	✓	✓	✗
Depth 1	No-worse vs Full	TITAN (Peng et al.)	✗	✗	✗	✗	✓
Depth 1	No-worse vs Full	TITAN+**Ours**	✗	✗	✗	✗	✓
Depth 1	No-worse vs Full	TITAN†	✗	✗	✗	✓	✓
Depth 1	No-worse vs Full	TITAN†+**Ours**	✓	✗	✗	✓	✓
Depth 2	ΔE vs Full ↓	**Ours**	−0.0632	−0.1081	−0.0765	−0.0583	−0.1241
Depth 2	ΔE vs Full ↓	Random	+0.2196	+0.5977	+0.5653	+0.8670	+3.6871
Depth 2	ΔE vs Full ↓	Global coeff.	−0.0965	−0.1624	−0.0270	−0.0491	+0.0889
Depth 2	ΔE vs Full ↓	TITAN (Peng et al.)	−0.0650	−0.2074	−0.0688	+0.0172	−0.1387
Depth 2	ΔE vs Full ↓	TITAN+**Ours**	−0.0603	−0.2072	−0.0736	+0.0162	−0.1369
Depth 2	ΔE vs Full ↓	TITAN†	−0.0819	−0.0621	−0.0283	−0.0596	+0.0998
Depth 2	ΔE vs Full ↓	TITAN†+**Ours**	+0.0120	−0.0710	−0.0400	−0.0239	+0.1058
Depth 2	Shot gain ↑	**Ours**	28.8%	32.8%	46.2%	52.6%	82.6%
Depth 2	Shot gain ↑	Random	29.9%	34.7%	47.5%	53.4%	83.9%
Depth 2	Shot gain ↑	Global coeff.	27.9%	33.2%	46.7%	52.0%	83.0%
Depth 2	Shot gain ↑	TITAN (Peng et al.)	53.2%	54.3%	51.2%	61.7%	78.7%
Depth 2	Shot gain ↑	TITAN+**Ours**	**70.5%**	**75.5%**	**80.9%**	**87.6%**	98.8%
Depth 2	Shot gain ↑	TITAN†	25.3%	40.2%	42.0%	48.2%	84.3%
Depth 2	Shot gain ↑	TITAN†+**Ours**	43.1%	64.2%	74.8%	80.9%	**99.4%**
Depth 2	No-worse vs Full	**Ours**	✓	✓	✓	✓	✓
Depth 2	No-worse vs Full	Random	✗	✗	✗	✗	✗
Depth 2	No-worse vs Full	Global coeff.	✓	✓	✓	✓	✗
Depth 2	No-worse vs Full	TITAN (Peng et al.)	✓	✓	✓	✗	✓
Depth 2	No-worse vs Full	TITAN+**Ours**	✓	✓	✓	✗	✓
Depth 2	No-worse vs Full	TITAN†	✓	✓	✓	✓	✗
Depth 2	No-worse vs Full	TITAN†+**Ours**	✗	✓	✓	✓	✗

Table 59: **Relevance sparsity under calibration-derived noise.** Metrics average four seeds at iterations 0, 14, and 27. Cross-mass is the noisy contribution mass retained by the clean top-mass mask.

System	Depth	Clean sparsity	Noisy sparsity	Mask Jaccard	Rank ρ	Cross-mass
H2 (4q)	1	47.3%	47.5%	0.996	0.986	99.9%
H2 (4q)	2	30.3%	30.7%	0.994	0.998	99.9%
HeH+ (4q)	1	49.9%	50.7%	0.981	0.996	99.7%
HeH+ (4q)	2	25.5%	26.0%	0.987	0.998	99.7%
H3+ (6q)	1	54.2%	54.6%	0.986	0.996	99.7%
H3+ (6q)	2	35.4%	36.1%	0.981	0.999	99.7%
H4 chain (8q)	1	68.4%	69.2%	0.967	0.996	99.6%
H4 chain (8q)	2	43.1%	44.5%	0.969	0.999	99.6%
LiH (12q)	1	82.3%	83.3%	0.938	0.995	99.7%
LiH (12q)	2	69.5%	72.4%	0.899	0.998	99.7%

Table 60: **Gate-level Aer spot-check under the FakeGuadalupeV2 snapshot** (IBM Quantum, 2026a;b). Each depth-1 contribution grid uses seed 97 and 4096 shots for every shifted QWC-group circuit. The selected-cell counts use $\tau = 0.995$. This is a calibrated Aer simulation, not a physical-QPU run.

System	Params.	QWC groups	Circuits	Clean selected	Noisy selected	Noisy sparsity	Jaccard / Rank ρ
H2 (4q)	8	5	80	21	34	15.0%	0.571 / 0.635
HeH+ (4q)	8	11	176	39	65	26.1%	0.529 / 0.441
H3+ (6q)	12	19	456	88	172	24.6%	0.477 / 0.335

Table 61: **Calibration-derived noisy Hamiltonian-panel transfer at $\tau = 0.997$.** This table repeats Table 39 under a fixed IBM `FakeGuadalupeV2`-derived Pauli-channel surrogate (Javadi-Abhari et al., 2024; IBM Quantum, 2026a;b). ΔE is paired against noisy Full-QWC at the same seed; final energies are evaluated noiselessly. TITAN† reuses the clean, cell-wise cost-matched thresholds without noise-specific tuning. Green/red ΔE and ✓/✗ follow the one-sided 1.6 mHa no-worse criterion; bold marks the closest mean ΔE to noisy Full, and bold-underlined shot gain marks the largest saving. This is calibrated simulation, not physical-QPU execution.

Testing Hamiltonian	Depth	Metric	**Ours**	Random	TITAN (Peng et al.)	TITAN+**Ours**	TITAN†	TITAN†+**Ours**
TFIM $h = 0.5$	Depth 1	ΔE vs noisy Full ↓	**-0.0030**	**-0.0112**	-0.0146	-0.0113	-0.0002	-0.0004
	Depth 1	Shot gain ↑	2.5%	2.3%	55.3%	**55.6%**	0.0%	2.1%
	Depth 1	Chem. acc.	✓	✓	✓	✓	✓	✓
	Depth 4	ΔE vs noisy Full ↓	+0.0203	+0.0104	+0.0033	+0.0040	**-0.0032**	+0.0110
	Depth 4	Shot gain ↑	2.1%	1.9%	55.2%	**57.1%**	0.0%	1.9%
	Depth 4	Chem. acc.	✗	✗	✗	✗	✓	✗
TFIM $h = 0.7$	Depth 1	ΔE vs noisy Full ↓	**+0.0010**	**-0.0061**	-0.0157	-0.0133	+0.0075	+0.0036
	Depth 1	Shot gain ↑	3.2%	3.0%	53.3%	**54.3%**	0.0%	2.8%
	Depth 1	Chem. acc.	✓	✓	✓	✓	✗	✗
	Depth 4	ΔE vs noisy Full ↓	+0.0063	+0.0123	**+0.0013**	+0.0028	**-0.0077**	**-0.0021**
	Depth 4	Shot gain ↑	1.9%	2.0%	47.6%	**49.8%**	0.0%	1.8%
	Depth 4	Chem. acc.	✗	✗	✓	✗	✓	✓
TFIM $h = 0.9$	Depth 1	ΔE vs noisy Full ↓	+0.1398	**+0.0032**	-0.0214	+0.0998	+0.0048	+0.1120
	Depth 1	Shot gain ↑	25.0%	25.0%	53.4%	**58.8%**	24.8%	38.5%
	Depth 1	Chem. acc.	✗	✗	✓	✗	✗	✗
	Depth 4	ΔE vs noisy Full ↓	+0.0909	+0.0181	**-0.0138**	+0.0887	**-0.0154**	+0.0357
	Depth 4	Shot gain ↑	6.2%	6.2%	42.9%	**46.2%**	4.5%	9.8%
	Depth 4	Chem. acc.	✗	✗	✓	✗	✓	✗
Heis. XXZ+X field	Depth 1	ΔE vs noisy Full ↓	**-0.1190**	+0.3294	+0.6037	+0.6111	+0.0398	**-0.0175**
	Depth 1	Shot gain ↑	16.7%	16.7%	52.2%	**53.7%**	18.9%	27.6%
	Depth 1	Chem. acc.	✓	✗	✗	✗	✗	✓
	Depth 4	ΔE vs noisy Full ↓	+0.3771	**+0.0446**	+0.2029	+0.1974	**-0.1167**	+0.7431
	Depth 4	Shot gain ↑	4.2%	4.2%	36.4%	**37.8%**	4.3%	7.5%
	Depth 4	Chem. acc.	✗	✗	✗	✗	✓	✗
Heis. J1–J2	Depth 1	ΔE vs noisy Full ↓	**-0.0907**	+0.4036	+0.5352	+0.5145	**-0.0779**	**-0.2654**
	Depth 1	Shot gain ↑	16.7%	16.7%	50.5%	**52.2%**	11.3%	22.8%
	Depth 1	Chem. acc.	✓	✗	✗	✗	✓	✓
	Depth 4	ΔE vs noisy Full ↓	+0.6422	**-0.0145**	+0.4860	+0.2510	+0.4632	+0.6444
	Depth 4	Shot gain ↑	4.2%	4.2%	34.2%	**36.1%**	4.1%	7.3%
	Depth 4	Chem. acc.	✗	✓	✗	✗	✗	✗
Heis. XXY+Z field	Depth 1	ΔE vs noisy Full ↓	**+0.0889**	+0.3424	+0.3749	+0.2889	**-0.1381**	**-0.2234**
	Depth 1	Shot gain ↑	16.7%	16.7%	49.3%	**52.1%**	14.1%	25.9%
	Depth 1	Chem. acc.	✗	✗	✗	✗	✓	✓
	Depth 4	ΔE vs noisy Full ↓	**-0.0345**	**-0.1825**	**-0.2512**	**-0.0173**	**-0.0943**	+0.1709
	Depth 4	Shot gain ↑	4.2%	4.2%	31.8%	**32.7%**	3.3%	7.0%
	Depth 4	Chem. acc.	✓	✓	✓	✓	✓	✗
JW-like parity spin	Depth 1	ΔE vs noisy Full ↓	**+0.0102**	+0.7538	+0.1011	+0.0976	**-0.0114**	+0.0106
	Depth 1	Shot gain ↑	36.4%	36.1%	56.2%	**77.1%**	29.6%	59.4%
	Depth 1	Chem. acc.	✗	✗	✗	✗	✓	✗
	Depth 4	ΔE vs noisy Full ↓	**-0.0024**	+0.0546	+0.0515	+0.0092	**-0.0346**	**-0.0106**
	Depth 4	Shot gain ↑	11.6%	11.6%	47.1%	**51.0%**	10.7%	19.3%
	Depth 4	Chem. acc.	✓	✗	✗	✗	✓	✓

L LONG-HORIZON CONVERGENCE AND FCI-ANCHORED VALIDATION

Table 62 reports the complete long-horizon benchmark referenced in Section 6. The exact-gradient Adam trajectories define attainable, seed-specific references for the fixed depth-3 ansatz; FCI is used only after optimization to quantify the remaining variational gap.

Table 62: **Long-horizon shot-noise-simulated VQE with an FCI-anchored diagnostic.** Gradients use Gaussian sampling from exact grouped-observable variances, not physical-QPU execution. All methods use the same depth-3 layered RY/RZ ansatz, three paired held-out seeds, 400 Adam steps, and 4096 shots per measured shifted QWC setting. Our method uses $\tau = 0.997$; Random receives its per-parameter group count, and TITAN (Peng et al., 2026) uses privileged exact-gradient ranking with the nearest Full-QWC parameter cost. The particle-number-penalized objective is $H + (N - N_e)^2$, while reported energies are those of H. FCI and exact statevector energies are diagnostics only and are never supplied to the finite-shot optimizer or stopping rule. *Exact floor* is the best exact-gradient Adam gap to FCI. *Parity* counts seeds reaching within 1.6 mHa of their exact-Adam reference at any checkpoint. No method reaches FCI chemical accuracy because the depth-3 variational floor remains 9.459–51.371 mHa above FCI. Green/red final-error and parity entries satisfy/violate the attainable-reference criterion. Within each molecule, bold marks the best final FCI gap, absolute final error, or shot saving; underline marks the corresponding runner-up. Shot savings remain black.

Molecule	Method	Exact floor (mHa)	Final FCI gap (mHa)	ΔE to exact ref. (mHa)	Shot saving	Parity
H2	Full-QWC	20.283 ± 0.303	20.672 ± 0.106	$+0.389 \pm 0.350$	$0.0 \pm 0.0\%$	3/3
	Random, same budget	20.283 ± 0.303	21.027 ± 0.222	$+0.744 \pm 0.305$	$26.3 \pm 1.7\%$	3/3
	TITAN, matched cost	20.283 ± 0.303	20.238 ± 0.718	-0.045 ± 0.415	$26.0 \pm 1.9\%$	3/3
	Ours	20.283 ± 0.303	20.673 ± 0.072	$+0.390 \pm 0.370$	$24.7 \pm 1.0\%$	3/3
HeH+	Full-QWC	9.459 ± 0.016	9.613 ± 0.159	$+0.154 \pm 0.158$	$0.0 \pm 0.0\%$	3/3
	Random, same budget	9.459 ± 0.016	9.663 ± 0.057	$+0.204 \pm 0.072$	$29.4 \pm 0.4\%$	3/3
	TITAN, matched cost	9.459 ± 0.016	9.446 ± 0.206	-0.014 ± 0.197	$30.5 \pm 0.2\%$	3/3
	Ours	9.459 ± 0.016	9.499 ± 0.111	$+0.040 \pm 0.103$	$29.8 \pm 0.9\%$	3/3
H3+	Full-QWC	11.939 ± 0.000	12.415 ± 0.187	$+0.477 \pm 0.187$	$0.0 \pm 0.0\%$	3/3
	Random, same budget	11.939 ± 0.000	13.127 ± 0.699	$+1.188 \pm 0.699$	$56.8 \pm 0.8\%$	3/3
	TITAN, matched cost	11.939 ± 0.000	12.220 ± 0.019	$+0.281 \pm 0.019$	$58.5 \pm 1.1\%$	3/3
	Ours	11.939 ± 0.000	12.274 ± 0.134	$+0.335 \pm 0.134$	$57.1 \pm 1.3\%$	3/3
H4 chain	Full-QWC	51.371 ± 0.000	51.981 ± 0.034	$+0.610 \pm 0.034$	$0.0 \pm 0.0\%$	3/3
	Random, same budget	51.371 ± 0.000	53.116 ± 0.897	$+1.745 \pm 0.897$	$76.7 \pm 0.3\%$	2/3
	TITAN, matched cost	51.371 ± 0.000	52.075 ± 0.161	$+0.704 \pm 0.161$	$77.6 \pm 0.1\%$	3/3
	Ours	51.371 ± 0.000	51.849 ± 0.100	$+0.479 \pm 0.100$	$76.9 \pm 0.7\%$	3/3
LiH	Full-QWC	18.415 ± 0.000	19.921 ± 0.437	$+1.507 \pm 0.437$	$0.0 \pm 0.0\%$	3/3
	Random, same budget	18.415 ± 0.000	728.571 ± 144.340	$+710.156 \pm 144.340$	$92.0 \pm 0.1\%$	0/3
	TITAN, matched cost	18.415 ± 0.000	19.673 ± 0.646	$+1.259 \pm 0.646$	$92.3 \pm 0.0\%$	3/3
	Ours	18.415 ± 0.000	20.098 ± 0.455	$+1.684 \pm 0.455$	$92.1 \pm 0.2\%$	3/3

M Implementation and Estimator Details

This appendix distinguishes the implemented learning objective, deployment inputs, and estimator used by each experiment. Classical predictor weights are fixed throughout quantum-parameter optimization. Updating circuit angles $\boldsymbol{\theta}$ with Adam is not fine-tuning the classical predictor or LLM.

M.1 Architecture and Feature Construction

The main predictor has two independent Transformer encoders, one for parameter tokens and one for QWC-group tokens. Each encoder has two pre-normalized layers, width 128, four attention heads, GELU activation, feed-forward width 512, and dropout 0.14. There is no cross-attention block: interaction occurs in the pairwise scorer described in Section 2. The scorer concatenates parameter and group embeddings, their elementwise product and absolute difference, and two pair features. Its LayerNorm–MLP has hidden widths 256 and 128, GELU nonlinearities, dropout, and a scalar output. A separate parameter-only LayerNorm–MLP with hidden width 128 predicts a sigmoid selected-group fraction. Variable-length examples are padded within a batch; padding masks exclude invalid groups and parameters from attention, normalization of row probabilities, and losses. The main saved state contains 984,454 trainable scalar parameters.

Actual inputs, not the entire historical feature interface. The saved interface has 35 parameter, 30 group, and two pair channels. Parameter descriptors encode normalized parameter, layer, and qubit indices, sinusoidal index features, rotation channel, and last-layer status. Group descriptors encode group index, coefficient ℓ_1 magnitude and its log/rank transforms, group/term counts, and Pauli-support fractions. Structural parameter/group summaries aggregate light-cone reachability. Pair channels are the fraction of group terms overlapping the parameter's light cone and the coefficient-weighted fraction of those terms. Current-state channels are exactly $[\sin\theta_j, \cos\theta_j, \tanh\theta_j, t/T]$; they are not a history of energies, gradients, previous masks, or optimizer moments. Historical cache and probe slots remain in the common interface but are not populated with oracle contributions in the direct relevance learner: token construction passes an empty cache and uses fixed action descriptors. Thus a feature name containing "history" does not imply access to past exact gradients. For each token type, training-set channel means and standard deviations standardize inputs; scales below 10^{-6} are replaced by one. Held-out examples use these saved statistics without refitting.

M.2 Exact Training Objective

Let $\mathcal{R}$ be valid rows with $\sum_k |C_{j,k}| > 10^{-10}$, $a_{j,k} = s_{j,k}/T_s$, and $q_j = \mathrm{softmax}(a_j)$ over valid groups. Here a row indexes a snapshot and parameter, so reductions below cover all valid rows in a batch. Let $m_{j,k}$ be the binary oracle top-mass mask at the *training* threshold τ_{train}, let $b_j = G_j^{-1}\sum_k m_{j,k}$, and let u_j be the auxiliary budget-head logit. The five terms in Eq. 6 are implemented as

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|\mathcal{R}|}\sum_{j\in\mathcal{R}}\sum_{k=1}^{G_j} w_{j,k}\log q_{j,k}, \tag{13}$$

$$\mathcal{L}_{\mathrm{mask}} = \frac{\sum_{j\in\mathcal{R}}\sum_{k=1}^{G_j}\mathrm{BCEWithLogits}(a_{j,k}, m_{j,k})}{\sum_{j\in\mathcal{R}} G_j}, \tag{14}$$

$$\mathcal{L}_{\mathrm{rank}} = \frac{\sum_{j\in\mathcal{R}}\sum_{k,l}(w_{j,k}-w_{j,l})_+\,\mathrm{softplus}(-(a_{j,k}-a_{j,l})/T_r)}{\max\{1, \sum_{j\in\mathcal{R}}\sum_{k,l}(w_{j,k}-w_{j,l})_+\}}, \tag{15}$$

$$\mathcal{L}_{\mathrm{budget}} = \frac{1}{|\mathcal{R}|}\sum_{j\in\mathcal{R}}\mathrm{SmoothL1}(\sigma(u_j), b_j), \tag{16}$$

$$\mathcal{L}_{\mathrm{entropy}} = -\frac{1}{|\mathcal{R}|}\sum_{j\in\mathcal{R}}\sum_{k} q_{j,k}\log q_{j,k}. \tag{17}$$

The entropy term has a positive coefficient and therefore penalizes diffuse predictions. BCE supervises the logits independently; it is not BCE applied to row-softmax probabilities. The budget

head is an auxiliary training loss for the main top-mass results, not their deployment rule. Historical count-head results instead round its predicted fraction times G_j and keep that many top-ranked groups. These two selection modes must not be conflated.

Table 63: Verified settings of the main molecular checkpoint and the separate depth-OOD checkpoint. Training and inference thresholds are different controls.

Component	Saved configuration / implementation
Training panel	Five molecular Hamiltonians; layered RY/RZ; depths 1–3
Training seeds	11, 23, 37, 41, 53, 67, 79, 83
Held-out seeds	97, 101, 107, 109
Label generation	18 exact-gradient Adam snapshots per cell; 2,160 snapshots
Angle initialization	Independent $\mathcal{N}(0, 0.35^2)$ angles
Quantum optimizer	Adam; learning rate 0.035; no gradient-norm clipping in the main 28-step runner
Classical training	260 epochs; batch size 8; AdamW; learning rate 3.5×10^{-4}; weight decay 2.5×10^{-4}
Classical gradient clip	Norm 1.0; initialization seed 123
Loss coefficients	$(\lambda_{\mathrm{m}}, \lambda_{\mathrm{r}}, \lambda_{\mathrm{b}}, \lambda_{\mathrm{e}}) = (0.15, 0.35, 0.60, 0.010)$
Temperatures	Training $T_s = 0.75$; ranking $T_r = 0.75$; inference 0.55
Thresholds	Main training mask 0.99; main inference top-mass 0.995 or 0.997
Checkpoint selection	Lowest epoch-average training loss; no held-out evaluation labels used for selection in this runner
Depth-OOD checkpoint	Separately trained on depths 1–4; 2,880 snapshots; training mask 0.995; evaluated at depths 5–7

M.3 Alternative Predictor Implementations

All four predictors receive the same standardized parameter, group, and pair features and the same training relevance labels in the main architecture comparison. The MLP, GNN, and ResNet-18 use separately trained checkpoints. Their saved settings agree with Table 63: width setting 128, dropout 0.14, 260 epochs, batch size eight, AdamW learning rate 3.5×10^{-4}, weight decay 2.5×10^{-4}, and initialization seed 123. The initialization and architecture differ, while the objective, normalization procedure, data split, and online protocol are fixed. Other transfer panels use their separately trained checkpoints.

Pairwise MLP. Independent parameter and group encoders each apply Linear–LayerNorm–GELU–Dropout–Linear–GELU with hidden width 128. The pair scorer concatenates the two embeddings, their product, their absolute difference, and the two pair features, then uses the same $514 \to 256 \to 128 \to 1$ head as the Transformer. The budget head uses LayerNorm and a $128 \to 128 \to 1$ MLP. This model has 224,902 parameters and no token-to-token attention.

Bipartite GNN. Parameter and group inputs are projected to width 128. A $2 \to 128 \to 128$ GELU MLP followed by a sigmoid produces a vector gate for every parameter–group edge. Two message-passing rounds alternately update parameter nodes from the mean gated valid-group embeddings, then group nodes from the updated valid-parameter embeddings. Each residual update applies LayerNorm and a $256 \to 128 \to 128$ GELU/dropout MLP to the node and its aggregated message. The pair and budget heads match the MLP above. Padding is excluded from message sums and denominators. This model has 407,942 parameters.

ResNet-18-style dense predictor. Linear projections map parameter, group, and pair inputs to 64 channels, which are added to form a $P \times G \times 64$ grid. A stride-one 3×3 convolution, one-group GroupNorm, and ReLU replace the image-classification stem. Four stages have two BasicBlocks each, channel widths $64, 128, 256, 512$, and first-block strides $1, 2, 2, 2$. Blocks contain two 3×3 convolutions, GroupNorm, ReLU, Dropout2d, and a projected skip when dimensions change. The final feature map is bilinearly interpolated to the input grid size with `align_corners=False`; GroupNorm–ReLU–1×1 convolution produces one score per pair. A separate parameter-only $64 \to 64 \to 1$ budget head is used. There is no ImageNet pretraining or fixed-resolution image resize. The model has 11,214,338 parameters.

Variable shapes and masking. Training pads each batch to its maximum parameter and group counts after feature standardization. Invalid output logits are masked and invalid rows are omitted from the loss. The Transformer masks padded attention keys; the GNN masks messages. ResNet convolutions and GroupNorm operate on the padded grid before output masking, so this implementation should not be assumed invariant to padding. Prediction uses one example at a time. These are architecture substitutions in the implemented common protocol, not independently optimized best configurations for each architecture.

Classical scaling. Let d be width and L_e the number of layers per encoder. With a fixed feed-forward multiplier, dense Transformer encoding requires

$$O\big(L_e[(P^2+G^2)d+(P+G)d^2]\big)$$

operations; the all-pairs MLP adds $O(PGd^2)$. Materialized attention and pair activations require $O(L_e(P^2+G^2)+PGd+L_e(P+G)d)$ memory for fixed head count, apart from weights and optimizer state. Sorting all rows for top-mass selection costs $O(PG\log G)$. These bounds describe inference after features and groups are constructed; they exclude offline simulation, grouping, and training. Efficient attention implementations can reduce materialized memory without removing the dense interaction cost.

M.4 Metric Definitions and Aggregation

For a molecule/ansatz/depth cell c and paired seed s, define

$$\Delta E_{c,s} = E_{c,s}^{\mathrm{method}}(T) - E_{c,s}^{\mathrm{Full}}(T), \qquad \overline{\Delta E}_c = \frac{1}{n_c}\sum_s \Delta E_{c,s}.$$

Energy entries are in Hartree (Ha), with 1 mHa $= 10^{-3}$ Ha. Energy $\pm$ values in the molecular, local-LLM, and frontier tables use population standard deviation over the recorded seeds (denominator n_c). Shot-gain uncertainties in Table 1 use sample standard deviation (denominator $n_c - 1$); its reused $K = 1$ rows retain that convention. The sparse-refresh and best-prompt shot-gain entries use population standard deviation. These values describe variation over the four paired seeds, not standard errors or confidence intervals.

Absolute gaps and table rankings. The main 5.67-mHa headline averages seed-level absolute gaps, $\frac{1}{|\mathcal{C}|}\sum_c \frac{1}{n_c}\sum_s |\Delta E_{c,s}|$. It is not the absolute value of the signed mean and need not equal the average of $|\overline{\Delta E}_c|$. The same seed-level convention is used for the frontier method summaries quoted in the main text. Bold/underline rankings and Random check/cross symbols in detailed tables compare unrounded cell means as specified by each caption. Coloring also uses unrounded means, so values printed identically can have different ranks or threshold outcomes. These are descriptive comparisons, not significance tests. *Best* ΔE in historical budget tables selects a depth-wise reported endpoint; it is not an average over depths.

Passing criteria and reference energies. *No worse* is one-sided: $\Delta E_{c,s} \le 0.0016$ Ha, so lower energy is accepted even outside the symmetric tolerance. Cell-level coloring applies this threshold to $\overline{\Delta E}_c$; seed counts apply it to individual trajectories. FCI error uses the separate exact ground-state reference. Legacy raw fields named `ca_vs_full` use the symmetric test $|\Delta E| \le 0.0016$ Ha; they must not be substituted for the one-sided no-worse counts. The long-horizon benchmark instead records whether a trajectory enters the 1.6-mHa neighborhood of its paired attainable exact-Adam reference at any checkpoint. Its final reference error, best observed reference attainment, and gap to FCI are separate quantities. The best exact-Adam energy found defines the empirical attainable reference for the chosen ansatz, initialization, and optimization schedule.

Measurement cost. For constant shots S per shifted selected group, $B = 2S\sum_t\sum_j |S_{j,t}|$ and $\mathrm{gain} = 100(1 - B_{\mathrm{method}}/B_{\mathrm{Full}})\%$. With adaptive allocation $n_{j,k,t}$, use $B = 2\sum_{t,j,k} n_{j,k,t}$. Reported aggregate gains average the per-trajectory ratios. A negative gain means more shots than the paired reference. Groups/parameter averages selected pairs over iterations and parameters; it is neither a count of distinct bases across the whole run nor the fraction of optimizer steps skipped. Refresh reuse holds the selection mask, not previously measured expectation values. Nominal counts for exact-expectation runs and sampled-gradient protocols are distinguished in Appendix N.1.

Feature-audit quantities. Captured mass for row j is $m_j = \sum_k M_{j,k}|C_{j,k}| / \sum_k |C_{j,k}|$ on nonzero rows; mean mass averages valid rows. The gradient cosine compares $\sum_k M_{j,k} C_{j,k}$ with the full gradient over all parameters. These quantities need not improve together because signed contributions can cancel. Jaccard compares the selected group sets against the full-feature reference action. A deletion effect in percentage points is $100(m^{\text{reference}} - m^{\text{deletion}})$ at matched snapshot and protocol; it is not a percentage relative improvement. Audit quality and token costs are conditional on the validity rules of the corresponding table, so context-window comparisons also report the number of valid cases.

M.5 Deterministic Truncation Bias

Condition on the current angles and predicted subset S_j. If measurements of retained groups are unbiased, the sparse estimator obeys

$$\mathbb{E}[\widehat{g}_j \mid \boldsymbol{\theta}, S_j] - g_j = -\sum_{k \notin S_j} C_{j,k}, \qquad |\mathbb{E}[\widehat{g}_j] - g_j| \le \sum_{k \notin S_j} |C_{j,k}|. \tag{18}$$

If the subset captures at least τ of the *exact* absolute mass, the last quantity is at most $(1 - \tau) \sum_k |C_{j,k}|$. This bound is conditional on exact retained mass; deployment applies the threshold to predicted mass. The implemented estimator uses deterministic truncation, whose bias is the omitted signed contribution above. The experiments evaluate the resulting closed-loop energy–cost trade-off. Offline exact-statevector label construction scales exponentially in qubit count, setting the computational challenge for larger training systems.

N Protocol Provenance and Numerical Diagnostics

N.1 Estimator Fidelity and Resource Accounting

The main 28-step molecular, transfer, and local/frontier-LLM ranking runners form gradients as $\sum_k M_{j,k}C_{j,k}$ using exact simulated expectations. Their shot counts are *nominal*: each retained parameter–group pair is charged two shifted settings times the specified shots per setting. The shot-count argument does not inject sampling noise in these runners. Randomness there comes from angle initialization and, for Random, sampled group identities, not from quantum shot sampling. Their energy differences therefore measure closed-loop truncation effects, not finite-shot robustness. Exact contributions may be calculated inside the simulator, but are not provided to our learned selector or to an LLM prompt. TITAN-style freezing controls additionally use privileged exact-gradient importance in simulation; their ranking information is stronger than a deployable zero-cost controller.

The separate shot-allocation panel samples gradient noise using exact grouped observable moments under the specified allocation, and the long-horizon benchmark uses a Gaussian finite-shot approximation with variance determined by those moments. These are simulated noisy estimators, not hardware measurements or exact sampling of every bitstring. The long-horizon benchmark uses 4096 shots per shifted selected group, depth 3, three paired seeds, 400 steps, and objective $H+(N-N_e)^2$, reporting energy of H. It is a distinct protocol, not a continuation of each 28-step run. It trains a separate depth-3 controller on 80-label-snapshot trajectories per cell, with training seeds 11/23/37 and validation seed 41; evaluation uses 97/101/107. Its model stays frozen during the 400-step quantum optimization. Its exact-Adam attainable references remain 9.459–51.371 mHa above FCI; Reference parity and absolute FCI error therefore quantify distinct aspects of the optimization outcome.

The calibration-derived Pauli-channel surrogate, finite-shot Aer grid spot check, and shot-noise VQE are also distinct experiments. The Aer check uses depth 1, seed 97, and 4096 shots per shifted group. Its 15–26% omitted cells describe *grid sparsity*, not realized learned-controller savings in a complete noisy online VQE. All reported noise studies use simulation. Online shot gain counts gradient measurement settings; it excludes offline label generation, classical training/inference, retrieval, compilation, energy diagnostics, and LLM token cost. Parameter counts, file sizes, tokens, and wall times are separate resource axes, not directly interchangeable costs. Historical dollar estimates use the original run-time accounting assumptions. Token traffic is the directly logged resource measure. The current scale comprises small molecular systems, four-seed short-horizon panels, and compact spin-model QWC partitions; larger systems require broader data generation and greater simulation resources.

Allocation controls and comparison scope. iCANS, gCANS, and Rosalin labels denote adapted allocation/operator-sampling policies inside a common Adam harness, with the caps, warm-up, and running moment estimates specified by the corresponding experiment. They do not use each paper's native optimizer or stopping rule: Adam, the horizon, and the experimental budget are shared controls. The panels evaluate allocation and operator-sampling policies, including their composition with learned QWC selection. The empirical baseline scope is this fixed-partition selection/allocation setting; related work separately discusses learned POVM and shadow-based measurement design.

Recorded allocation settings. The main allocation panel uses five molecules, depths 1–2, seeds 97/101/107/109, 28 Adam steps at learning rate 0.035, and no gradient-norm clipping. Nominal shots are $1024b$ per shifted group, with $b \in \{0.25, 0.50, 1.00\}$. The learned mask uses the main saved checkpoint, inference temperature 0.55, and top-mass threshold 0.995. The allocation state starts with zero gradient and coefficient-based variance prior $\max(\|h_{\mathcal{G}_k}\|_1, 10^{-4})^2$. After observations, gradient and group-variance forecasts use exponential averaging with decay 0.9. Variance observations are simulated from the exact grouped moments using a chi-square sample with $n-1$ degrees of freedom; the gradients use a Gaussian moment approximation.

The adapted CANS policies use $L = 10$, $\eta = 0.035$, and $a = 2\eta L/(2-\eta L)$. Their unrounded per-pair requests are

$$s_j^{\mathrm{iCANS}} = a\frac{\sum_k \widehat{v}_{j,k}}{\widehat{g}_j^2 + 10^{-8}}, \qquad s_j^{\mathrm{gCANS}} = a\frac{\sigma_j \sum_i \sigma_i}{\sum_i \widehat{g}_i^2 + 10^{-8}}, \quad \sigma_j^2 = \sum_{k\in S_j} \widehat{v}_{j,k}.$$

Requests are multiplied by b, rounded upward, and clipped to $[\lceil 64b \rceil, \lceil 4096b \rceil]$. The first uninitialized step uses the lower bound. These are shots per selected pair, so the row cost also scales with its number of retained groups.

Variance-aware allocation keeps each row's nominal total and distributes it in proportion to predicted standard deviations, with a minimum of two shots per selected group and stable largest-remainder integer rounding. The combined learned-mask/gCANS control additionally uses this within-mask reallocation. Rosalin uses the adapted iCANS row request times the full group count, multinomial group sampling proportional to coefficient ℓ_1 magnitude, and inverse-probability correction. All methods then take the same Adam update.

Small estimated gradients can drive the CANS request toward its cap; therefore shared nominal scale does not enforce equal realized shot cost, and negative gains are possible. This explains how overspending arises in this implementation without establishing why a native algorithm would perform poorly. The stored run records one configuration for these constants, not a dedicated hyperparameter search for each native optimizer. The comparison supports composition under the stated harness.

N.2 Threshold and Checkpoint Provenance

Table 9 is a historical aggregate, not a single-checkpoint threshold sweep. Its 0.98 and 0.99 rows use separately trained checkpoints with different oracle training-mask thresholds and the historical count-head selection mode. Its 0.95 row is an earlier trained operating point. Only the 0.995 and 0.997 rows use the same saved $\tau_{\mathrm{train}} = 0.99$ checkpoint with top-mass post-processing. The main molecular baseline panel also loads that checkpoint. The Transformer checkpoint used for LLM cardinalities and native attribution has identical model-state and normalization values; only its serialized artifact/configuration hash differs. In contrast, Table 10 reuses one separately trained depth-OOD checkpoint over the displayed inference thresholds. Thus the nonmonotone group counts between historical 0.99 and 0.995 rows are not a violation of fixed-prediction top-mass nesting. Even for a fixed checkpoint, different masks change subsequent angles, so closed-loop average group counts and final energies need not be monotone in the threshold. A checkpoint transferred within a named study should not be described as one universal checkpoint across separately trained depth-, molecule-, and Hamiltonian-family panels. "Best ΔE" columns are selected descriptive endpoints, not estimates of average reliability; depth-wise and seed-wise results remain the relevant evidence for that question.

N.3 Train-Only Retrieval for Frontier Models

The retrieval bank is restricted to the same molecule, ansatz, depth, and parameter ID, using training seeds only. For a parameter row, distances are squared Euclidean distances between $[\sin\theta_j, \cos\theta_j, \tanh\theta_j, t/T]$, standardized by the training bank's per-channel standard deviation (values below 10^{-4} are replaced by one). Invalid zero-mass rows are excluded and ties use a stable sort. The prompt receives two explicit nearest-neighbor records and a summary over the eight-neighbor pool, including average training relevance and selection frequencies, with a limited list of salient group signals. It is not an unbounded dump of the full training dataset. Retrieval uses training relevance labels, never the current held-out contribution matrix. Both use exact offline relevance supervision, but the frontier bank uses 20 snapshots per training cell (2,400 total), versus 18 (2,160 total) for the main learned checkpoint. Data access is therefore not matched in snapshot count, and in-context examples are not equivalent to gradient-based training. Unlike the local fixed-cardinality tests, frontier models choose both identities and counts: the request explicitly removes each parameter's k field and accepts either a select or exclude list with at least one measured group. This differs in both data access and count policy from the local tests, so the comparison evaluates the combined model and data interface. The requested 0.995 mass is an objective informed by retrieved training labels; the validator checks output structure, range, and cardinality. The four-refresh schedule is shared with our $K = 7$ control, but budgets are not matched; comparisons are accuracy–cost operating points. This evaluates data-exposed frozen selection within an external harness, not a fine-tuned LLM or autonomous code-executing agent.

N.4 Decision Traces, Prompt Identity, and Reason Tags

The action and reason tags in trace Box (4) come from the audit panel's `explain_tags` prompt. The full-feature reference and deletions in Box (7) come from `baseline`, `no_coeff`, `no_lightcone`, and `no_theta_time`. All four decision traces use H_3^+, depth 2, seed 97, iteration 0, parameter 14, and cardinality six. The decision-trace state is initialized with $\theta_j \sim \mathcal{N}(0, 0.35^2)$, giving $\theta_{14} = +0.1067889170$ rad. The 24 circuit parameters and 19 QWC groups describe problem dimensions, not tokenizer counts. Local contribution shares, mean captured mass, and gradient cosine are evaluated on this same state. The integrated-gradient studies teacher-force the reference action using their separately serialized attribution inputs (Appendix N.5). The prompt-specific groups are:

Model	Explanation-prompt action	Reference action / IG target
Qwen3-8B	$[0, 1, 2, 3, 4, 5]$	$[0, 1, 2, 3, 4, 5]$
Llama-3.1-8B	$[7, 8, 9, 10, 11, 12]$	$[4, 5, 6, 7, 8, 9]$
DeepSeek-R1-8B	$[0, 2, 3, 4, 6, 7]$	$[2, 5, 9, 6, 0, 7]$
GPT-OSS-20B	$[18, 0, 17, 16, 15, 14]$	$[0, 1, 2, 6, 7, 3]$

The raw Qwen explanation tags are `lightcone`, `theta`, and `circuit_location`; Llama emits `coefficient`, `theta`, `iteration`; DeepSeek emits `coefficient`, `lightcone`; GPT-OSS emits `coefficient`. Box (3) reproduces these logged tags. Each action's exact contribution shares are computed *after* inference, outside the model, solely for evaluation. An intervention difference must be subtracted from the reference-prompt score, not from a different explanation-prompt score. Llama's explanation action captures 1.797%; its reference action captures 1.442%, and deleting coefficients raises that local reference mass to 94.843% (+93.401 percentage points). Over valid parameters, mean mass instead falls from 0.483 to 0.395, distinguishing the local intervention effect from aggregate selection quality.

Dependence, usefulness, and faithfulness are distinct. A changed mask establishes input sensitivity in the reference protocol. A positive captured-mass drop means the deleted feature was useful there; a negative drop means it was harmful there, not that it was unused. Unchanged outputs can reflect redundancy or feature interactions. Reason tags and deletion effects are linked by snapshot while retaining their distinct prompt identities: `explain_tags` produces the reported tags, and reference-prompt interventions measure feature utility. The verdicts in the drawings are interpreted through this distinction. These traces record observable interaction and externally measured sensitivity. Local LLMs run frozen text inference with code-execution tools disabled; `think=false` specifies the direct-output configuration used in this comparison.

Protocol-specific output contracts. The rank-only/repaired local protocol in Appendix J emits a ranking. The clean protocol instead uses `selections.p14 = {mode, group_ids}`, emitting either the selected IDs or their shorter excluded complement at exact cardinality. The audit's optional tags use the controlled vocabulary `coefficient`, `lightcone`, `theta`, `circuit_location`, `iteration`, and `group_support`, with one to three unique tags per row. The clean validator checks integer type (excluding booleans), range, uniqueness, mode, and length; schema feedback permits retries, but no coefficient completion repairs IDs. Persistent failure invokes logged Full-QWC safety fallback and charges all groups. The audit additionally varies output contracts and optionally returns `reason_codes`; the best-prompt evaluation tests its selected prompt online. The example `{parameter_id, mode, group_ids, reason_codes}` in a trace is a normalized logged record, not the complete literal clean-protocol wire schema. The best prompt was selected on the fixed audit panel, which includes seed 97, also present in the online evaluation. It is the best tested configuration on that tuning panel; the shared seed links prompt selection to the corresponding online evaluation cells. The `pair_nonzero_rows` key names the column dictionary, whereas `pair_nonzero` holds each parameter's rows; the different names are intentional. These are frozen LLM inference protocols within our external Python harness, not learned tool-use agents.

Clean system instruction and explanation extension. The clean instruction, also used by the domain-prompt audit baseline, is:

```
You are a measurement-selection controller for variational
quantum eigensolvers.  For every requested circuit
parameter, follow its fixed output mode and emit exactly the
requested number of distinct QWC group IDs.  In select mode,
emit the groups most relevant to that parameter's energy
gradient, ordered from most to least relevant.  In exclude
mode, emit the groups least relevant to that gradient,
ordered from least to most relevant; every unlisted
candidate will be measured.  Use Hamiltonian coefficient,
circuit location, theta, iteration, and parameter--group
light-cone features jointly.  A globally large coefficient
does not imply equal relevance to every parameter.  Return
only the JSON object required by the schema.  Never omit a
requested parameter and never invent an ID. The exact
cardinality is mandatory.
```

For the explanation condition the appended instruction is:

```
For each parameter, also emit one to three reason_codes
chosen only from ['coefficient', 'lightcone', 'theta',
'circuit_location', 'iteration', 'group_support'].  These are
concise evidence tags, not chain-of-thought.
```

The domain reference uses a 32,768-token context window and the compressed select/exclude contract. A row with requested count $k_j \leq G - k_j$ uses select mode and emits k_j IDs; otherwise it emits the $G - k_j$ excluded IDs. For the shared trace with $G = 19$ and $k_{14} = 6$, the explanation response has the following wire structure (shown with the accepted Qwen action):

```
{"selections": {
  "p14": {
    "mode": "select",
    "group_ids": [0, 1, 2, 3, 4, 5],
    "reason_codes": ["lightcone", "theta", "circuit_location"]
  }
}}
```

Without explanation tags the same reference contract omits `reason_codes`. The request also specifies each row's `selected_count`, `output_mode`, and `emit_count`. The returned object is validated against those values, range and uniqueness constraints, and the complete set of requested parameter keys.

What the context ablation varies. The context ablation changes the configured context-window capacity to 8,192, 16,384, or 32,768 tokens. This is a capacity setting, not a promise that every request fills that many tokens. Its separate prompt variants use a minimal instruction or add two training demonstrations; output variants emit a direct selected list or a complete ranking. Deletion and shuffling conditions hold the Full-QWC snapshot and requested cardinalities fixed. Explanation tags are an additional condition. Reported differences compare configurations under this harness; context capacity, protocol success, and valid-case mask quality describe distinct parts of controller performance.

Observable computation boundary. The models serialize a selection as text/JSON; Python constructs features, retrieves training records where enabled, validates outputs, applies masks, and performs the VQE update. Code-execution and external tool calls are not available to the evaluated local models, and frontier tool use is disabled in the saved CLI protocol. These controls specify the observable interaction and the division of work between model and harness. The frontier GPT runs record low reasoning effort; the local direct-output tests record `think=false`. Conclusions concern those settings.

N.5 Integrated-Gradient Numerical Reliability

The four LLM attribution prompts use the serialized H_3^+, depth-2, seed-97, iteration-0 state with initialization scale 0.12 ($\theta_{14} \simeq +0.0366$ rad) and the fixed reference group lists above. The decision traces use their recorded scale-0.35 state; each analysis evaluates its own state and score while sharing the task and target IDs. For the LLM figures, we fix the observed action and teacher-force the JSON prefix for each emitted ID. The scalar score is the first-token log probability conditioned on grammar-valid ID prefixes, not the probability of the entire freely generated mask. Only marked numeric feature-token embeddings are interpolated. The input is tokenized once: the baseline replaces these embeddings with repeated embeddings of the token for numeric zero, preserving sequence length, positions, attention mask, syntax, instructions, and the fixed output prefix. It does not retokenize a string of shorter numerical values. This is a synthetic embedding-space baseline, not a physically valid alternative VQE state. Attribution sums are aggregated by labeled spans. Absolute-mass normalization summarizes the distribution of attribution magnitude across input families; its denominator is total absolute attribution, whereas energy accuracy is evaluated separately.

The Ollama/GGUF generation runtime is not differentiable in this setup; gradients come from matched Hugging Face model families under the quantization specified by each figure. Quantization, kernels, and routed experts can change scores and continuations. Our controller instead uses its native checkpoint, standardized scalar features, a training-mean baseline, and 1024 quadrature nodes. Therefore cross-model percentages are not calibrated on a common functional or numerical scale; each panel describes its own target score and baseline.

For a scored group, let $D = f(x) - f(x_0)$ and $R = D - \sum_i \mathrm{IG}_i$. We report both absolute and relative error: small $|R|$ alone is insufficient when $|D|$ is near zero. The table below is computed from the saved metadata, not from a new experiment. Qwen's first score has a 1.61% residual relative to its score change; several later near-saturated continuation scores have substantially larger relative errors. The Llama, DeepSeek, and GPT-OSS entries likewise show substantial finite-quadrature residuals. Their plotted family shares are finite-quadrature estimates under the specified scoring functions. The native controller uses 1024 nodes and has maximum relative residual 0.46%; its 12.16% current-state attribution identifies a resolved dependence of its group scores on optimization-state inputs. The LLM panels use different scoring functions and embedding baselines, so the family percentages describe within-model attribution profiles. Closed-loop energy and shot cost provide the common cross-model performance measures.

Table 64: Completeness diagnostics for the six displayed group scores. The last column counts scores with $|R| > |D|$; these estimates fail even a 100% relative completeness check. Residuals accompany the finite-quadrature estimates for each model.

Model	Nodes	Max $\lvert R\rvert$	Max $\lvert R\rvert/\lvert D\rvert$	$\lvert R\rvert > \lvert D\rvert$
Qwen3-8B	32	0.1224	8.56	4/6
Llama-3.1-8B	32	1.1491	93.33	3/6
DeepSeek-R1-8B	32	0.4373	4.63	2/6
GPT-OSS-20B	32	1.4140	1.40	2/6
Our controller	1024	0.0126	0.0046	0/6

N.6 Fixed-Snapshot Attribution Stability Audit

We test whether the feature-family ordering in Figure 6 persists across numerical integration settings. This completed follow-up evaluates the same four serialized H_3^+ attribution cases described in Appendix N.5: depth 2, seed 97, iteration 0, parameter 14, and initialization scale 0.12. Each model retains its archived prompt, token positions, reference group IDs, teacher-forced JSON prefix, score function, and numeric-zero embedding baseline. These are numerical repeats of fixed cases, not additional physical snapshots or online VQE runs.

Controlled refinement. We use 32, 64, 128, and 256 Gauss–Legendre nodes with float64 accumulation. All four models are evaluated under their original attribution runtime. For Qwen, Llama, and DeepSeek, a separate FP32-compute diagnostic preserves the NF4 weight bytes while promoting floating-point layers and quantized matrix-multiplication compute to FP32. This is a separate arithmetic regime, not an unquantized model or a bitwise-identical scoring implementation. GPT-OSS retains its native MXFP4 implementation and hard expert routing. The audit thus contains 28 model–precision–quadrature configurations, each scoring six fixed group IDs. No model is trained, no new action is generated, and the original plotted assets are retained.

Aggregation and ranking. For output group g, let I_{gs} be the signed IG summed over the tokens of numeric input span s. Its feature-family share is

$$A_{gF} = 100\,\frac{\sum_{s\in F}|I_{gs}|}{\sum_{s}|I_{gs}|}, \qquad \overline{A}_F = \frac{1}{6}\sum_{g} A_{gF}. \tag{19}$$

The family ranking uses $\overline{A}_F$. Percentages therefore summarize absolute span attribution, not selection probability, causal effect size, or energy contribution. Structure includes circuit location and Pauli support; budget/action covers controller-contract fields. Trajectory denotes the serialized current-angle and normalized-step features, rather than a complete history of measured gradients.

Table 65: Feature-family ranking and trajectory-share stability in the fixed-snapshot audit. S: circuit/support; C: coefficient; L: light cone; B: budget/action; T: trajectory. Each model's mean ranking is identical across its tested settings. Shares are averaged over six scored group IDs; the range spans all tested node counts and arithmetic regimes, not a confidence interval. FP32 is a compute diagnostic with unchanged NF4 weights.

Model	Stable mean ranking	Trajectory at 256 nodes (%)		All-setting range (%)
		Original	FP32	
Qwen3-8B	$S > C > L > B > T$	0.77	1.04	0.59–1.06
Llama-3.1-8B	$S > C > L > B > T$	0.95	0.76	0.67–0.95
DeepSeek-R1-8B	$S > C > L > B > T$	0.99	1.01	0.96–1.01
GPT-OSS-20B	$S > L > C > B > T$	2.49	–	2.49–2.71

Consistent low trajectory attribution. Trajectory ranks fifth among the five families for every scored group in every configuration: 168/168 output–configuration combinations. The model-specific orderings and trajectory-share ranges are reported in Table 65. Under the original runtime, increasing from 128 to 256 nodes changes these means by only 0.024, 0.165, 0.007, and 0.037 percentage points. These counts describe numerical robustness within the fixed cases, rather than independent-sample statistical evidence.

Implication for frozen selection. Despite receiving the full input feature set, frozen LLMs do not automatically make substantial use of trajectory information without targeted intervention in these audited cases: static physical inputs dominate their attribution profiles. This pattern persists across the tested quadrature and precision settings. It complements the feature-deletion and online evaluations in Tables 56 and 55, which assess changes in decisions and closed-loop performance. The consistently low within-model trajectory ranking, rather than an exact cross-model percentage ratio, is the central result of this refinement.

Numerical checks and their relation to ranking stability. The audit predeclares a per-output completeness tolerance $|R| \leq 0.001 + 0.05|D|$, with D and R defined in Appendix N.5. Between successive node counts, it also checks maximum family-cell change ≤ 1 percentage point, maximum trajectory-cell change ≤ 0.5 percentage points, and minimum per-output family-rank Spearman correlation ≥ 0.9. Refinement can stop after at least 128 nodes only when two consecutive levels meet all conditions; otherwise it continues through 256 nodes. All seven model–runtime pairs reach 256 nodes in this audit.

Table 66: Final numerical diagnostics, kept separate from the stable mean rankings in Table 65. Residual and completeness pass counts use 256 nodes. Changes and correlations compare 128 with 256 nodes: Δ_F is the largest absolute family-cell change, Δ_T the largest trajectory-cell change (both in percentage points), and $\rho_{\min}$ the minimum per-output rank correlation. Pass counts apply the combined absolute/relative completeness tolerance, not a relative-error-only threshold.

Model	Compute	Max $\lvert R\rvert$	Complete scores	Δ_F	Δ_T	$\rho_{\min}$
Qwen3-8B	Original	1.8269	0/6	3.296	0.205	0.90
Qwen3-8B	FP32	3.4425	2/6	2.551	0.528	0.90
Llama-3.1-8B	Original	0.9630	0/6	5.381	0.568	1.00
Llama-3.1-8B	FP32	0.3602	2/6	2.144	0.134	1.00
DeepSeek-R1-8B	Original	0.2188	1/6	1.276	0.047	0.90
DeepSeek-R1-8B	FP32	0.0222	4/6	0.812	0.013	1.00
GPT-OSS-20B	Original	1.4116	0/6	1.838	0.352	1.00

DeepSeek with FP32 compute meets the family-stability conditions at 256 nodes, with a maximum family-cell change of 0.812 percentage points and identical per-output family ranks. Its completeness tolerance is met by four of six scores. Across all seven model–runtime pairs, the number of complete scores ranges from zero to four, so the full conjunction of numerical criteria is not met. The other pairs retain larger family-cell changes even when their mean ordering and low trajectory shares persist. Thus the audit reports ranking robustness and completeness as separate properties; the percentages remain finite-quadrature estimates. In particular, normalizing absolute attribution to 100% is not a correction for residual error. No post-hoc rescaling is applied.