QuPID: Quantum Parameter-Efficient Input-Dependent Retrieval Adaptation for Medical RAG

Anonymous authors
Paper under double-blind review

Abstract

Fidelity-based quantum retrieval ranks candidates by the fidelity between query and archive states. Applying a shared input-independent unitary after fixed state encoding leaves that fidelity unchanged, so training the circuit cannot alter the ranking. Quantum parameter-efficient input-dependent retrieval adaptation (QuPID) repairs this by making the circuit input-dependent through data re-uploading and by comparing measurement readouts, vectors of local Pauli expectations, rather than states. The result is a small readout for adapting frozen image features to a local archive with limited data: training simulates the circuit classically, and inference runs on a GPU with fixed learned parameters. We characterize the class as a structured factorization of input-modulated quadratic feature maps, bound the frequency support of its re-uploading channel, and give a parameter-count generalization bound that motivates its small budget. Under a shared frozen backbone and a label-free protocol, QuPID's 60 parameters give higher precision-at-5 (P@5) on ChestX-ray14 and MURA than frozen medical encoders, and than adapters and low-rank adaptation (LoRA) with up to 5.25 million trainable parameters. On ChestX-ray14, the P@5 gain over the frozen encoder is $+0.116$, the lead over retuned adapters is widest at 512 adaptation examples ($+0.040$), and the full-budget margin over an equally compact classical rotation-plane head is $+0.023$ with a 95% interval excluding zero. Medical imaging is the primary testbed; the pattern recurs on two non-medical benchmarks, in report generation, and under simulated gate noise and finite-shot readout.

1 Introduction

Can a 60-parameter readout adapt a frozen retriever as well as adapters with millions of parameters? Medical case retrieval makes the question concrete. A pathology occupies a small part of the image, and once the encoder pools it, shared anatomy dominates, so a 'Normal' study and an 'Early Pneumonia' study can be nearly collinear in a frozen feature space (Cheng et al., 2022; Kienitz et al., 2022). Retrieval-augmented generation (RAG) built on that geometry (Jiang et al., 2025; Wang et al., 2025a; Fan et al., 2024) provides the generator with diagnostically inconsistent evidence, resulting in hallucination (Huang et al., 2025; Tivnan et al., 2024). Data protection rules can keep images inside the hospital (Yang et al., 2025), while the local archive may offer few adaptation examples. Our aim is a model that adapts retrieval geometry with few trainable parameters while retaining competitive ranking quality. Medical imaging is our primary setting; industrial inspection and fine-grained recognition provide additional tests of adaptation at a small parameter budget (Section 4.1).

We use a variational quantum circuit to share trainable parameters across measurement features (Benedetti et al., 2019; Roh & Kim, 2025). Here *quantum* denotes a circuit-defined hypothesis class and not a computational-advantage claim, in the sense argued for by Schuld & Killoran (2022): every circuit in this paper is simulated classically. For a diagnostic control motivated by state-based representations (Lisnichenko & Protasov, 2023; Jaiswal et al., 2021), we amplitude-encode query and database features, apply one shared trainable circuit $U(\boldsymbol{\theta})$, and rank by fidelity $|\langle\psi_q|\psi_p\rangle|^2$. The shared-unitary identity rules out adaptation here. Unitarity cancels the circuit exactly, so the score equals the squared inner product of the frozen features for every $\boldsymbol{\theta}$ and the contrastive gradient is identically zero (Proposition 1). Minibatch variation does not imply a circuit-dependent ranking.

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.

The proof also identifies how to break this invariance, since the score depends on the two states only through an overlap that any shared input-independent isometry preserves. Either the map after encoding depends on the input, or the comparison passes through a non-unitary readout, and QuPID (quantum parameter-efficient input-dependent retrieval adaptation) does both: *data re-uploading* layers (Pérez-Salinas et al., 2020) make the transformation input-dependent, and similarity compares *measurement readouts*, vectors of local Pauli expectations, rather than states.

This construction gives a trainable similarity function; its practical value depends on the geometry produced by the shared parameterization. We claim neither dimensional expansion, since at $n_q = \log_2 d$ qubits the state space has exactly the input dimension, nor hardness of simulation, since a shallow circuit with local observables is the regime this literature expects to be classically simulable. Our contribution is a structured retrieval feature map: 60 shared parameters jointly control $M = 40$ measurement outputs through norm-preserving gates, local Pauli observables and input-dependent re-uploading. The analysis characterizes the resulting quadratic forms and a controlled re-uploading spectrum (Proposition 2). Proposition 3 gives a generic parameter-count bound for bounded smooth classes on a compact parameter domain. We compare frozen medical encoders, scaled adapters, low-rank adaptation (LoRA) and parameter-matched controls on ChestX-ray14 and MURA with multi-seed statistics. With 60 parameters, QuPID attains higher precision-at-5 (P@5) than adapters and LoRA with up to 5.25M parameters, and its lead over retuned adapters is widest with 512 adaptation examples; an equally compact classical rotation-plane head narrows the gap to a full-budget margin whose paired interval excludes zero. Non-medical retrieval, report generation, simulated noise and finite-shot readout assess its scope, alongside cost and metric-specific losses.

2 Related Work

Retrieval-augmented generation in medicine. RAG reduces hallucination in medical large vision-language models (LVLMs) by conditioning generation on context retrieved with pre-trained vision encoders (Wu et al., 2025; Lin et al., 2025; Bain et al., 2021; Liu et al., 2021), but flat Euclidean embeddings entangle clinically distinct, visually similar cases (Kienitz et al., 2022; Wang et al., 2025b). Domain-specialized encoders such as BiomedCLIP (Zhang et al., 2025) and MedCLIP (Wang et al., 2022) are trained centrally and do not address data-local adaptation; we use them as frozen references and as a QuPID backbone.

Parameter-efficient adaptation. Adapters and low-rank updates such as LoRA (Hu et al., 2022) are standard parameter-efficient alternatives. We compare them at scaled budgets alongside classical heads near 60 parameters. Quantum circuits have entered parameter-efficient adaptation on the weight side, where Quantum-PEFT builds a full-rank update from a Pauli-structured unitary (Koike-Akino et al., 2025), quantum parameter adaptation generates LoRA weights from a circuit during training (Liu et al., 2025) and QuanTA factorizes a high-rank update into circuit-shaped tensors (Chen et al., 2024b), and on the activation side, where quantum-amplitude embedded adaptation compresses attention activations into amplitudes and transforms them with a circuit (Roh & Kim, 2025). None of these methods places the circuit in a retriever's similarity function, so none faces the degeneracy of Section 3.1. In QuPID the data are amplitude-encoded and re-uploaded, and the circuit defines a retrieval feature map sharing parameters across its measurement outputs. Like the weight-side methods, it uses no quantum hardware (Appendix A.6). Random features (Rahimi & Recht, 2007) approximate kernels and supply a compact classical nonlinear control.

Quantum machine learning for representation. Amplitude embedding encodes d-dimensional vectors in $\log_2 d$ qubits (Schuld & Killoran, 2019; Gonzalez-Conde et al., 2024), parameterized circuits define trainable transformations on the encoded states (Benedetti et al., 2019; Haug et al., 2021), and data re-uploading interleaves encoding and trainable layers, which controls the Fourier spectrum accessible to circuit outputs (Pérez-Salinas et al., 2020; Schuld et al., 2021). Reading out classical features through measurements connects our design to projected quantum kernels (Huang et al., 2021) and to shadow-style models that learn from local observables (Li et al., 2021). QuPID builds on measurement-based learning but targets label-free cosine retrieval: its shared local readout is trained through a contrastive objective, and Section 3.1 shows why state overlap cannot adapt under an input-independent shared unitary; medical retrieval applications remain sparse (Wei et al., 2023). This literature also supplies the standard by which our claims should be judged: variational models can be approximated by classical surrogates that sample their Fourier

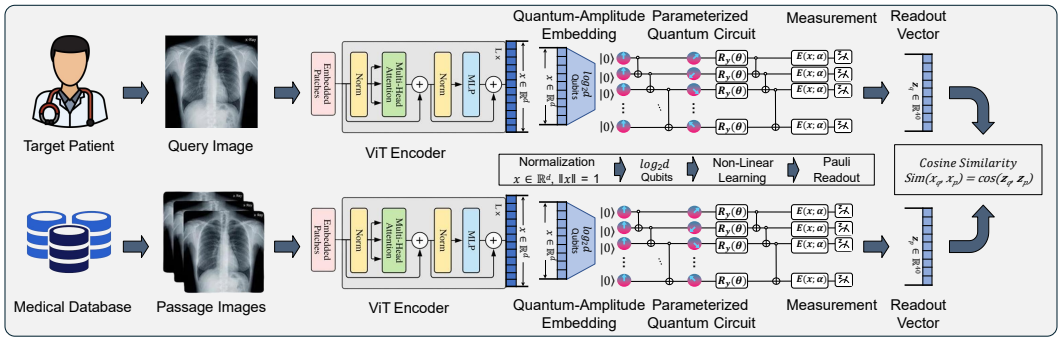


Figure 1: QuPID overview. A frozen vision transformer (ViT) encodes all images; each feature is amplitude-encoded into $\log_2 d$ qubits and processed by a shared circuit (trainable rotations, CNOT ring, re-uploading gates $E(\mathbf{x}; \boldsymbol{\alpha})$) with all 60 trainable parameters. Retrieval compares Pauli readouts $\mathbf{z} \in \mathbb{R}^{40}$ by cosine similarity, not quantum states (Section 3.1); archive readouts are cached.

spectrum (Landman et al., 2023), tuned classical baselines outperform quantum models on most benchmarks studied so far (Bowles et al., 2024), and shallow local circuits are, where understood, classically simulable (Cerezo et al., 2025). We adopt all three as constraints: no claim rests on hard simulation, every comparison shares the backbone, and parameter-matched, budget-retuned and wall-clock-matched controls are reported alongside.

3 Method

3.1 Setup and a degeneracy pitfall in quantum retrieval

Let $\mathcal{D} = \{(\mathbf{x}_i, r_i)\}_{i=1}^N$ be a hospital archive of images with diagnostic reports, let a frozen ViT produce features $\mathbf{h} = \mathrm{ViT}(\mathbf{x}) \in \mathbb{R}^d$, $d = 1024$, and let conventional RAG rank by cosine similarity to the query feature. We seek a lightweight transformation of this geometry, learned from unlabeled local data, under which pathologically consistent pairs receive higher scores than pairs that merely share anatomy, which Figure 1 summarizes end to end.

Amplitude embedding maps the normalized feature $\hat{\mathbf{h}}$ to an n_q-qubit state $|\psi_{\mathrm{in}}(\mathbf{x})\rangle = \sum_{k=0}^{d-1} \hat{h}_k |k\rangle$ with $n_q = \log_2 d$ (Figure 7; background in Appendix A, implementation in Appendix D). The natural variational design applies a trainable circuit $U(\boldsymbol{\theta})$ to both query and database states and ranks by fidelity:

$$\mathrm{Sim}_{\mathrm{fid}}(\mathbf{x}_q, \mathbf{x}_p; \boldsymbol{\theta}) = \left| \langle \psi_{\mathrm{in}}(\mathbf{x}_q) | U(\boldsymbol{\theta})^\dagger U(\boldsymbol{\theta}) | \psi_{\mathrm{in}}(\mathbf{x}_p) \rangle \right|^2. \tag{1}$$

Proposition 1 (Degeneracy of shared-unitary fidelity retrieval)**.** *For any input-independent unitary $U(\boldsymbol{\theta})$, equation 1 satisfies $\mathrm{Sim}_{\mathrm{fid}}(\mathbf{x}_q, \mathbf{x}_p; \boldsymbol{\theta}) = |\langle \psi_{\mathrm{in}}(\mathbf{x}_q) | \psi_{\mathrm{in}}(\mathbf{x}_p) \rangle|^2$ for all $\boldsymbol{\theta}$. Thus equation 1 induces squared-cosine ranking on the amplitude-encoded frozen features. Any loss whose only dependence on $\boldsymbol{\theta}$ is through these similarities is constant in $\boldsymbol{\theta}$ and has identically zero gradient.*

The proof is immediate from $U^\dagger U = I$ (Appendix B, verified numerically). The result extends to every similarity that depends on the states only through their overlap, including the Fubini–Study distance, the trace distance and monotone functions of fidelity (Corollary 1). Changing the metric therefore cannot resolve the degeneracy. Fidelity yields squared-cosine rather than cosine ranking in general; the two orders coincide in our evaluated rank ranges because the query-gallery cosines there are nonnegative (Appendix B). This is a diagnostic identity, not a new expressivity theorem or evidence that an established retriever has this defect; Appendix B.4 tests the deliberately constructed control. Proposition 1 also identifies the required modification: the circuit must depend on the input beyond the initial encoding, or the similarity must pass through a non-unitary map such as measurement. QuPID does both.

3.2 AMPLITUDE ENCODING WITH DATA RE-UPLOADING

QuPID processes the encoded state with L blocks, each consisting of a trainable rotation layer, an entangling layer, and a data re-uploading layer:

$$|\psi(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\alpha})\rangle = \prod_{l=1}^{L}\Big[E\big(\mathbf{x};\boldsymbol{\alpha}^{(l)}\big)\,U_{\mathrm{ent}}\,U_{\mathrm{rot}}\big(\boldsymbol{\theta}^{(l)}\big)\Big]\,|\psi_{\mathrm{in}}(\mathbf{x})\rangle, \tag{2}$$

where $U_{\mathrm{rot}}(\boldsymbol{\theta}^{(l)}) = \bigotimes_{j=1}^{n_q} R_Y(\theta_j^{(l)})$ applies one trainable R_Y rotation per qubit, U_{ent} is a ring of CNOT gates introducing correlations between qubits, and the re-uploading layer $E(\mathbf{x};\boldsymbol{\alpha}^{(l)}) = \bigotimes_{j=1}^{n_q} R_Y\big(\alpha_j^{(l)} s_j(\mathbf{x})\big)$ re-injects the input through fixed block-pooled projections of the feature vector, scaled by trainable coefficients $\alpha_j^{(l)}$. The fixed projections carry a per-qubit dyadic factor, $s_j(\mathbf{x}) = 2^{j-1}\bar{s}_j(\mathbf{x})$ with $\bar{s}_j$ the pooled block mean, so that the encoding matches the premise of the spectral analysis in Section 3.3. All gate matrices are real, so the state remains real-valued. With n_q qubits and L blocks the trainable budget is $2n_qL$; the default ($n_q = 10$, $L = 3$) has 60 parameters. The measurement readout of Section 3.3 alone escapes the degeneracy of Proposition 1; re-uploading additionally enriches the function class (Proposition 2); Section 4.2 separates the two.

3.3 MEASUREMENT READOUT AND SIMILARITY

Rather than comparing states, QuPID compares measurement statistics. We define the readout map

$$\mathbf{z}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\alpha}) = \Big(\langle\psi|Z_j|\psi\rangle,\ \langle\psi|X_j|\psi\rangle,\ \langle\psi|Z_jZ_{j+1}|\psi\rangle,\ \langle\psi|X_jX_{j+1}|\psi\rangle\Big)_{j=1}^{n_q} \in \mathbb{R}^M, \tag{3}$$

with indices taken cyclically, giving $M = 4n_q = 40$ components from single-qubit and nearest-neighbor Pauli observables, and writing $|\psi\rangle \equiv |\psi(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\alpha})\rangle$ for brevity. Retrieval similarity is the cosine between readouts, $\mathrm{Sim}_{\mathrm{qt}}(\mathbf{x}_q,\mathbf{x}_p) = \mathbf{z}(\mathbf{x}_q)^\top\mathbf{z}(\mathbf{x}_p)/(\|\mathbf{z}(\mathbf{x}_q)\|\,\|\mathbf{z}(\mathbf{x}_p)\|)$. Each component of $\mathbf{z}$ has the form $\langle\psi_{\mathrm{in}}|V^\dagger PV|\psi_{\mathrm{in}}\rangle$ for the input-dependent circuit V and a Pauli observable P; we suppress the dependence of $\mathrm{Sim}_{\mathrm{qt}}$ on $(\boldsymbol{\theta},\boldsymbol{\alpha})$. Without re-uploading each component is exactly a quadratic form $\hat{\mathbf{h}}^\top A_m(\boldsymbol{\theta})\hat{\mathbf{h}}$; with re-uploading, $A_m(\boldsymbol{\theta},\boldsymbol{\alpha},\mathbf{x})$ is additionally modulated by trigonometric functions of linear projections of the input, so the readout consists of *input-modulated* quadratic forms. The M readouts share one circuit, and each realized A_m is orthogonally similar to a Pauli observable, hence full-rank with ±1 spectrum. An unrestricted rank-r factorization uses $O(dr)$ parameters per feature and is full-rank only at $r = d$. Here 60 parameters jointly specify M structured forms; full rank alone does not establish an expressivity advantage. The design shares a small parameter vector across correlated, input-modulated quadratic features. The circuit preserves the state norm, while re-uploading and measurement let the similarity change: this enables training of the design. Proposition 3 uses bounded outputs and parameter derivatives.

Re-uploading also controls the readout's functional richness. We specialize the encoding-dependent Fourier analysis of Schuld et al. (2021) to a clamped-state scalar probe of our re-uploading channel (Pérez-Salinas et al., 2020); Appendix B gives the derivation.

Proposition 2 (Spectral expressivity of the re-uploading channel)**.** *Clamp the amplitude-encoded state at a reference input and let a scalar probe t drive the re-uploaded angles $s_j = 2^{j-1}t$. For fixed $\boldsymbol{\alpha}$ every readout component of equation 3 is a finite trigonometric polynomial in t whose frequencies are signed sums of the rates $\{\alpha_j^{(l)}2^{j-1}\}_{j,l}$. At $\alpha_j^{(l)} = 1$ those frequencies are integers and every ω in the support satisfies $|\omega| \le L(2^{n_q}-1)$.*

At $\boldsymbol{\alpha} = 1$ this conservative bandwidth bound grows exponentially in n_q; local readout gives a tighter bound, and boundary attainment is not asserted (Appendix B). Three qualifications delimit the scope of this result. The statement concerns a controlled probe of the re-uploading channel, not the end-to-end map, since the amplitude-encoded state varies with the same input. Reaching a frequency is not controlling it: only $2n_qL$ coefficients are independently adjustable, and classical product-form features also reach exponential support. No lower bound against classical surrogates follows, so Section 4 tests parameter-matched controls rather than inferring superiority from the spectrum.

3.4 LABEL-FREE CONTRASTIVE ADAPTATION

Site adaptation must proceed without expert labels, so we adopt SimCLR-style contrastive learning (Chen et al., 2020): two pathology-preserving augmentations of an image form a positive pair, in-batch samples act as negatives, and for views i with positives i^+,

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\alpha}) = -\frac{1}{2B} \sum_{i=1}^{2B} \log \frac{\exp\big(\mathrm{Sim}_{\mathrm{qt}}(\mathbf{x}_i, \mathbf{x}_{i^+})/\tau\big)}{\sum_{j \neq i} \exp\big(\mathrm{Sim}_{\mathrm{qt}}(\mathbf{x}_i, \mathbf{x}_j)/\tau\big)}, \tag{4}$$

with temperature $\tau = 0.07$ and conservative, lesion-preserving augmentations (Appendix D). Every gate generator in equation 2 has two eigenvalues, so gate-angle derivatives obey exact parameter shift; derivatives of re-uploading scales include $s_j(\mathbf{x})$, with the classical chain rule through cosine and loss (Appendix D). In-batch negatives can include false negatives when two patients share a pathology; Appendix D quantifies the collision rate.

Proposition 3 (Parameter-count generalization bound)**.** *Fix $R, c > 0$, let $p = 2n_qL$, and let $\Theta_{R,c}$ collect the $(\boldsymbol{\theta}, \boldsymbol{\alpha}) \in [-R, R]^p$ for which $\|\mathbf{z}(\mathbf{x})\| \geq c$ almost surely. On $\Theta_{R,c}$ the loss equation 4 is bounded and L_{lip}-Lipschitz in the parameter ∞-norm, with L_{lip} polynomial in p, M, $1/c$, $1/\tau$, and $\sup_{j,\mathbf{x}} |s_j(\mathbf{x})|$. With probability $1 - \delta$ over n adaptation samples, every empirical risk minimizer over this class has generalization gap $O\big(\sqrt{(\,p \log(1 + nRL_{\mathrm{lip}}) + \log(1/\delta)\,)/n}\big)$. The leading rate is $\sqrt{p/n}$, with M and the input scales entering only logarithmically, and the same conclusion holds for any p-parameter class of the same boundedness and smoothness.*

The proof (Appendix B.3) is a covering argument enabled by the structure: Pauli expectations are bounded and parameter-shift bounds every coordinate derivative. The rate is generic to any p-parameter class with the same smoothness, norm floor and compact domain; it bounds contrastive risk rather than retrieval accuracy; and the reported training does not impose that box constraint. We therefore interpret it as motivating a capacity hypothesis, rather than providing a guarantee for the reported runs. Section 4 evaluates the retrieval quality attained by this parameterization at a small trainable budget. Larger adapters and equally compact classical heads provide distinct comparisons.

Neither proposition establishes trainability (Gil-Fuster et al., 2025), and circuit gradients can vanish exponentially in depth and qubit count (McClean et al., 2018). QuPID avoids that regime by design: $L \leq 6$, n_q fixed at 10 by the feature dimension, and local one- and two-body observables, a setting associated with milder shallow-circuit decay (Cerezo et al., 2021; Basheer et al., 2025). Those guarantees do not cover entangled amplitude-encoded inputs, so Figure 11 measures gradient variance from one to twelve layers (four times the default) as a diagnostic, not an asymptotic rate.

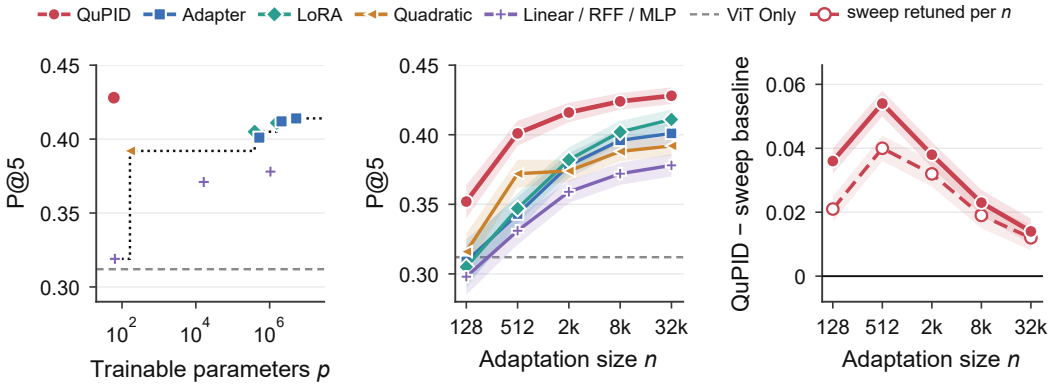
Deployment. After classical circuit-simulation training, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are fixed. GPU inference computes each query readout, compares it with cached archive readouts, and passes the top-k reports to a frozen medical LVLM. The query-dependent re-uploading gates are evaluated on the GPU using 1024 amplitudes; no quantum processor is required. On-site execution reduces data movement, not formal privacy risk (Appendix A.5).

4 EXPERIMENTS

Setup. Retrieval is evaluated on *ChestX-ray14* (Wang et al., 2017) and *MURA* (Rajpurkar et al., 2018), generation on *IU X-Ray* (Demner-Fushman et al., 2016) and *MIMIC-CXR* (Johnson et al., 2019), with patient-disjoint medical splits. Appendix C gives split counts, exclusions and metric definitions. Every adapted method starts from the same ViT-L/16 and shares augmentations, temperature and schedule: adapters, Linear, Tiny MLP, RFF and Quadratic use frozen features, while LoRA updates attention projections with the base weights frozen. Section 4.2 adds a 60-parameter rotation-plane control. Training uses the label-free objective of equation 4; labels enter only post-hoc evaluation, never model selection or tuning, with default hyperparameters fixed on a disjoint pilot subset; the separate retuning sweep selects by contrastive validation loss. The five-seed means overlap (Figure 13); † marks Holm-corrected query-bootstrap significance conditional on fitted runs. Appendix M.2 separately reports normal-approximation intervals against the rotation-plane head across adaptation draws, initialization and patients.

Table 1: Retrieval under the shared label-free protocol (mean±std, 5 seeds); frozen encoders are deterministic. Within ViT-L, **bold**/underline mark best/second; tints mark 1st, 2nd, 3rd. † denotes conditional query-bootstrap significance versus the strongest baseline in this table (Holm-corrected, $p < 0.05$; Section 4). The final medical-backbone block is a separate comparison.

Method	ChestX-ray14			MURA			#Params
	P@5	MAP@10	NDCG@10	P@5	MAP@10	NDCG@10	
ViT Only	0.312	0.278	0.307	0.481	0.451	0.475	0
BiomedCLIP	0.352	0.315	0.348	0.508	0.476	0.501	0
MedCLIP	0.341	0.303	0.334	0.494	0.463	0.488	0
Linear Head	0.378±.008	0.341±.007	0.368±.008	0.549±.007	0.513±.008	0.541±.007	1.05M
Adapter	0.401±.007	0.368±.006	0.396±.007	0.573±.006	0.546±.007	0.567±.006	525K
Adapter-L	0.412±.006	0.372±.007	0.401±.006	0.581±.006	0.556±.006	0.580±.007	2.10M
Adapter-XL	0.414±.009	0.387±.008	0.414±.008	0.584±.008	0.562±.008	0.584±.008	5.25M
LoRA (r=4)	0.405±.008	0.370±.008	0.399±.008	0.576±.007	0.549±.007	0.571±.007	393K
LoRA (r=16)	0.411±.007	0.376±.007	0.405±.007	0.582±.006	0.557±.007	0.579±.006	1.57M
Tiny MLP	0.371±.010	0.337±.009	0.365±.010	0.545±.009	0.512±.009	0.538±.009	16.4K
RFF	0.319±.005	0.283±.005	0.312±.005	0.489±.005	0.458±.005	0.483±.005	64
Quadratic	0.392±.009	0.355±.009	0.383±.009	0.561±.008	0.532±.008	0.554±.008	164
QuPID	**0.428**†±.006	**0.399**†±.006	**0.425**†±.006	**0.601**†±.006	**0.579**†±.006	**0.599**†±.006	**60**
BiomedCLIP+QuPID	0.438±.006	0.409±.006	0.435±.006	0.612±.006	0.590±.006	0.610±.006	54

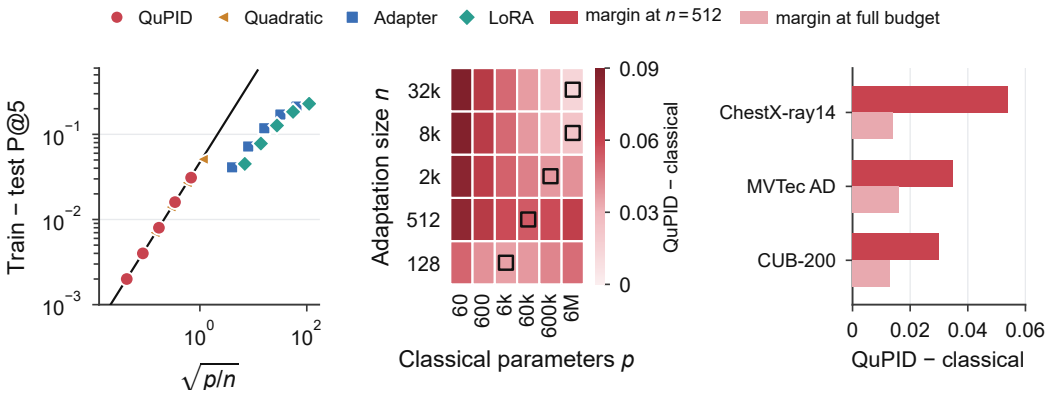


(a) Parameter budget. (b) Adaptation size. (c) Margin in the retuning sweep.

Figure 2: Low-data retrieval (ChestX-ray14, 5 seeds; bands one std). The dotted step in (a) is the Pareto envelope of the plotted classical modules; (b) plots the adaptation-size sweep. Panel (c) uses the fixed and retuned comparisons of Table 22; it excludes the additional controls in Table 23. Proposition 3 does not predict these retrieval margins.

4.1 Main retrieval results

Table 1 and Figure 2(a) evaluate the accuracy attained at each trainable budget; Figure 8 (Appendix E) gives metric-specific differences. With 60 parameters, QuPID exceeds every adapter and LoRA configuration in Table 1 on both datasets. ChestX-ray14 MRR is lower. Relative to the frozen encoder, QuPID improves P@5 by +0.116 and +0.120; its measured margins over a five-million-parameter adapter are +0.014 and +0.017. These results demonstrate useful retrieval adaptation with four to five orders of magnitude fewer trainable parameters than the adapters. In Table 6, trained readout recovers 0.100 of the 0.116 gain; adding re-uploading contributes 0.016 while increasing the budget from 30 to 60. This ablation changes both structure and capacity. An untrained circuit achieves half of the gain, and the naive fidelity design matches the frozen baseline, as Proposition 1 and the observed nonnegative query-gallery cosines require. ChestX-ray14 label-set Jaccard describes neighborhood structure; on MURA it equals P@10 by definition (Table 10).



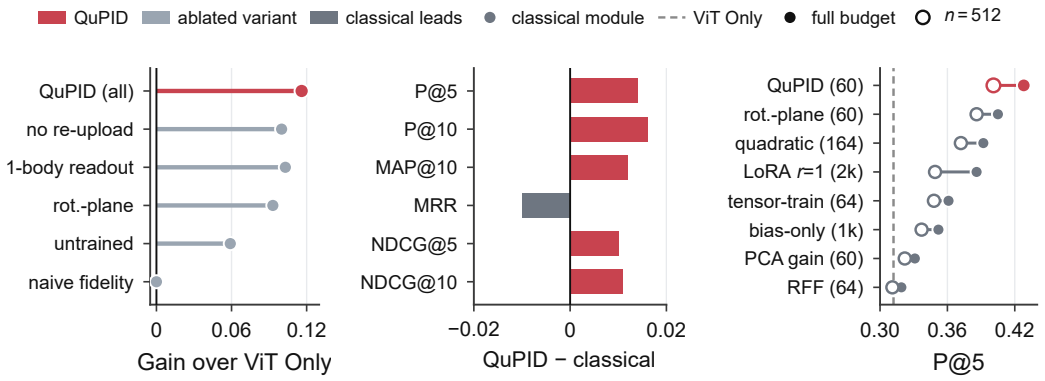
(a) Gap against $\sqrt{p/n}$. (b) Margin across p and n. (c) Margin in three domains.

Figure 3: Locating the advantage (ChestX-ray14 unless noted, 5 seeds). The solid line in (a) is an unfitted slope-one guide, and the squares in (b) mark the adapter budget with the best test score at each n. Absolute numbers are in Appendices E and J.

Where the parameter budget stops paying. The adaptation-size sweep evaluates the compact map under changing data budgets. Proposition 3 motivates restricting capacity but does not predict an accuracy ordering. Figures 2 and 3 compare the plotted families; Table 23 supplies the closest matched control. Below roughly one thousand examples adapters degrade toward, and partially below, the frozen baseline. At 512, QuPID reaches 0.401 P@5 versus 0.386 for the default 60-parameter rotation-plane head, the strongest control in Table 23, a $+0.015$ margin. The same-control full-budget margin is $+0.023$. Full-to-512 drops are 0.027 for QuPID and 0.019 for the control: higher accuracy, but greater degradation. Appendix J.1 suggests why larger heads fall behind: the adapter's training P@5 exceeds its held-out P@5 by 0.041 at the full budget and 0.171 at 512, against 0.002 and 0.016 for QuPID. The rotation-plane sweep in Appendix J.3 measures accuracy and train–test gaps at all five adaptation sizes. Figure 3(b) sweeps parameter budget and data budget jointly: the classical parameter budget achieving the highest test score increases with adaptation size, and our margin is always smallest at those budgets. Retuning the configurations in Table 22 reduces their 512-example margin from $+0.054$ to $+0.040$ (Figure 2(c)). These are comparisons within that sweep, not margins over all classical controls. Validation-selected rotation-plane tuning reduces the margins to $+0.007$ at 512 and $+0.014$ at full budget (Appendix M.1). Paired 95% intervals exclude zero at the full budget and not at 512, for the default control ($[-0.0112, +0.0412]$ at 512) and for the tuned control (Table 24).

Does the effect leave radiology. The readout can process any frozen feature vector satisfying the encoding requirements of Section 3. We test whether its utility extends beyond radiology. Figure 3(c) repeats the protocol on two non-medical benchmarks that share the structure of Section 1. *MVTec AD* (Bergmann et al., 2019) evaluates object-category retrieval in industrial imagery; *CUB-200* (Wah et al., 2011) evaluates species retrieval. These relevance definitions test transfer beyond radiology; the MVTec result does not directly measure defect localization. With backbone, readout and hyperparameters unchanged, the same pattern appears in both, the low-data margin roughly twice the full-budget one (Table 5). These comparisons report mean accuracy for the included baselines; they do not isolate the size of the discriminative signal as the cause.

Does better retrieval reach the generator. Retrievers are adapted label-free on each generation dataset's training split, which also serves as the archive and is patient-disjoint from all queries; decoding is greedy with a fixed prompt, so seed variation stems from the retriever alone. Table 7 (Appendix E) shows QuPID improving lexical-overlap metrics for both generators and datasets, while Adapter retains a small CIDEr advantage. Because n-gram overlap cannot certify clinical correctness, Table 8 probes factual content: CheXbert-F1 improves over Adapter by $0.018/0.017$ and RadGraph-F1 by $0.011/0.012$ for LLaVA-Med/CheXagent. The comparison is meaningful only if generation responds to retrieval, so Appendix K brackets the pipeline between random and label-oracle context; QuPID recovers 61% of this clinical-proxy interval and Adapter 49% (Figure 16).



(a) Ablations and controls. (b) Per-metric margin. (c) Classical capacity ladder.

Figure 4: Mechanism (ChestX-ray14, 5 seeds; P@5 except (b)). Gains in (a) are over the frozen encoder; (b) uses the strongest included comparator; (c) gives parameter counts, filled at full budget and hollow at n=512. Full sets are in Appendices E, H and J.

4.2 Isolating the mechanism

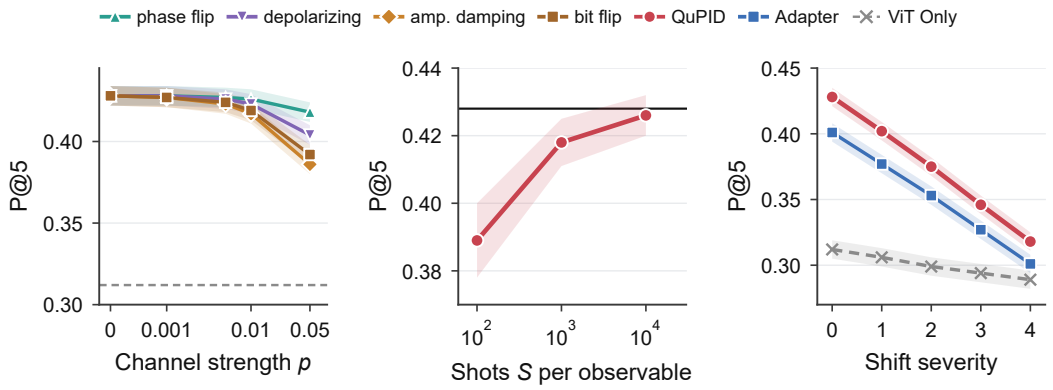
The comparisons above assess the utility of a compact map. We next examine its design choices and equally small classical alternatives. These controls distinguish the comparison with larger adapters from comparisons between parameterizations at a similar budget. The ten-qubit real circuit is simulated exactly with 1024-dimensional linear algebra, so these comparisons evaluate representational choices within classical simulation (Gil-Fuster et al., 2025).

The circuit couples 40 measurement features through one norm-preserving transformation with 60 trainable parameters. For each input, the associated quadratic forms share an orthogonal conjugation of fixed Pauli observables. Re-uploading makes this shared transformation input-dependent, tying how the features can adapt. Under classical GPU execution, the circuit thus specifies an architectural constraint whose retrieval utility is evaluated by the parameter-matched controls.

Figure 4(a) removes one component at a time and Table 6 (Appendix E) gives the full set. Depth saturates beyond $L = 3$; single-qubit and pairwise readouts both contribute. The rotation-plane head matches our budget and isometry but uses input-independent rotations and rank-one squared-projection forms. Its comparison with input-modulated full-rank Pauli forms evaluates the combined design, without separating modulation, readout rank and entanglement (Appendix D). The same parameter budget can therefore support different feature families. A variational readout admits a Fourier description and can be approximated classically by sampling frequencies from its spectrum (Landman et al., 2023), but Proposition 2 bounds our bandwidth without providing a lower bound on the cost of such a surrogate. RFF and the four-projection Quadratic head are restricted controls, not exhaustive dequantizations; matching parameter count does not match retained information. A random Fourier map recovers 0.007 of the 0.116 that separates the readout from the frozen encoder and a trained quadratic head 0.080, leaving 0.036 over Quadratic; the stronger rotation-plane head narrows this margin to 0.023. Appendix J.4 builds that surrogate from the spectrum of Proposition 2 at 5,000× our budget: its full-budget P@5 trails QuPID by 0.021 (paired 95% interval $[+0.013, +0.029]$) and by 0.039 at 512 examples. Figure 4(c) compares classical modules at or near our budget under their default configurations (Appendix J.3); the readout has the highest mean P@5 at both adaptation sizes, with the isometric control closest. The larger-adapter comparison establishes parameter efficiency at competitive accuracy; equal-budget controls assess the specific structure chosen for that budget.

Figure 4(b) shows the per-metric boundary of the claim. Appendix H varies qubit count, entangling topology, readout width, temperature and feature noise, and every axis saturates at or near the default, so the operating point is a plateau rather than a tuned optimum. It also reports per-seed spread and a zero-shot cross-site margin over the adapter that stays positive, falling from 0.027 to 0.017.

Computational cost. Table 9 (Appendix E) reports the tradeoff: QuPID requires 1.60 times the per-epoch training time of a lightweight adapter, because statevector simulation scales with $2^{n_q} = d$. Inference (7.6 against 7.1 ms per query) and peak adaptation memory (3,582 against 3,728 MB)



(a) Gate noise at evaluation. (b) Finite-shot readout. (c) Acquisition shift.

Figure 5: The idealizations tested (ChestX-ray14, P@5, 5 seeds; bands one standard deviation). Severity in (c) runs from the ChestX-ray14 acquisition at 0 to the CheXpert (Irvin et al., 2019) acquisition at 4. The strength axis in (a) is a symmetric log; in (b) the horizontal rule marks exact expectations and the points are measured finite-shot results.

are comparable, since archive readouts are cached and each query requires one circuit evaluation. Parameter efficiency is not computational efficiency. At matched adaptation time, Table 29 preserves the ordering among its three baselines; Appendix M.1 separately evaluates rotation-plane tuning with an explicit trial budget. The GPU deployment of Section 3.4 incurs this classical cost; Appendix A.5 describes its fixed-parameter forward pass.

Testing the idealizations. GPU deployment uses exact statevector readouts. Figure 5 tests acquisition shift and, separately, noise and sampling in a hypothetical quantum-hardware realization; the latter is not required for our GPU inference path. Gate noise degrades retrieval gradually, and the channel ordering follows the cancellation argument of Appendix I.1. The evaluated Pauli channels multiply Pauli components at the channel location by observable-dependent factors, without an additive offset. Cosine removes a common scaling factor; amplitude damping additionally biases expectations, consistent with its larger measured degradation. Training through the channel recovers part of the loss but never the noiseless level (Figure 14 and Tables 18 and 19, Appendix I). Finite-shot readout is within 0.002 P@5 of exact expectations at $S = 10^4$ (Table 20); this is an empirical result, separate from the sufficient ranking guarantee in equation 13. Figure 5(c) and Appendix L report positive acquisition-shift margins and less sequential forgetting than either adapter. Figure 17 indexes stored cache size; patient exclusion determines the eligible gallery used for ranking. The scope is narrow: these simulations omit crosstalk, coherent errors and readout error. They show that the margin survives the modeled perturbations, not that it would survive on hardware today.

5 Limitations and Conclusion

Five limitations bound our claims. First, training uses classical circuit simulation and inference runs on a GPU; no computational quantum advantage is claimed. Second, the method operates on pooled backbone features, so information averaged away by the frozen [CLS] aggregation cannot be recreated downstream, and patch-token variants are future work. Third, amplitude encoding requires strict L2 normalization and power-of-two dimensions, so a 768-dimensional backbone needs zero-padding to 1024. Fourth, the retriever is trained in isolation from the generator, generation metrics lack clinician assessment, and data-local adaptation is not formal privacy protection. Fifth, the margin over the 60-parameter rotation-plane head is positive at every adaptation size; paired intervals exclude zero at the full budget but not at 512 examples, including after validation-selected tuning.

QuPID adapts a frozen retriever through 60 parameters shared across 40 input-dependent measurement features, and this structure exceeds adapters and LoRA with up to 5.25M parameters in P@5. Its P@5 gain over the frozen encoder is $+0.116$, its lead over retuned adapters is widest at 512 examples ($+0.040$), and its full-budget margin over an equally compact classical head is $+0.023$. Circuit-defined measurement features are thus a practical route to parameter-efficient retrieval adaptation under classical GPU execution.

ETHICS STATEMENT

All datasets used in this work are publicly distributed, and the medical datasets are de-identified. ChestX-ray14 and MIMIC-CXR were collected under IRB approval by their curators. MIMIC-CXR access follows the PhysioNet credentialed data use agreement, MURA and CheXpert are distributed by Stanford under their research use agreements, and IU X-Ray is distributed de-identified through Open-i. MVTec AD is used under its CC BY-NC-SA 4.0 license, and CUB-200 under the terms of its Caltech distribution. The proposed system is intended to assist, not replace, clinical decision-making: any deployment must operate under physician supervision, undergo applicable regulatory review, and include mechanisms for flagging low-confidence retrievals. We explicitly do not claim privacy-preserving learning; data-local adaptation reduces data movement but does not defend against parameter-level attacks such as membership inference on the adapted module.

REPRODUCIBILITY STATEMENT

Proofs of Propositions 1, 2 and 3 are given in Appendix B. Implementation details, hyperparameters, augmentation settings, and seed handling are specified in Appendix D, and all datasets are available from their public distributors under the terms listed in the Ethics Statement. Code is available at `https://anonymous.4open.science/r/qupid-iclr/`.

AI USE STATEMENT

We used generative AI tools to edit and restructure the text, verify and format references, write figure and reference code, and assist with proof revision and with drafting interpretive sentences, which the authors checked against the measured results. We did not use them to generate data or to produce any reported experimental value; translation and qualitative analysis are not applicable. The authors reviewed all AI-assisted content and take full responsibility for the paper.

REFERENCES

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1728–1738, Virtual, Oct. 2021.

Afrad Basheer, Yuan Feng, Christopher Ferrie, Sanjiang Li, and Hakop Pashayan. On the trainability and classical simulability of learning matrix product states variationally. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, pp. 15498–15506, Philadelphia, PA, USA, Feb. 2025.

Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, Oct. 2019.

Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M Sohaib Alam, Guillermo Alonso-Linaje, B AkashNarayanan, Ali Asadi, et al. PennyLane: Automatic differentiation of hybrid quantum-classical computations. arXiv:1811.04968, Nov. 2018.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD: A comprehensive real-world dataset for unsupervised anomaly detection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592, Long Beach, CA, USA, Jun. 2019.

Jannis Born, Filip Skogh, Kahn Rhrissorrakrai, Filippo Utro, Nico Wagner, and Aleksandros Sobczyk. Quantum doubly stochastic transformers. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 70224–70254, San Diego, CA, USA, Dec. 2025.

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

Matthias C. Caro, Elies Gil-Fuster, Johannes Jakob Meyer, Jens Eisert, and Ryan Sweke. Encoding-dependent generalization bounds for parametrized quantum circuits. *Quantum*, 5:582, Nov. 2021.

M. Cerezo, Akira Sone, Tyler Volkoff, Lukasz Cincio, and Patrick J. Coles. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, 12(1):1791, Mar. 2021.

M. Cerezo, Martin Larocca, Diego García-Martín, Nahuel L. Diaz, Paolo Braccia, Enrico Fontana, Manuel S. Rudolph, Pablo Bermejo, Aroosa Ijaz, Supanut Thanasilp, Eric R. Anschuetz, and Zoë Holmes. Does provable absence of barren plateaus imply classical simulability? *Nature Communications*, 16(1):7907, Aug. 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. International Conference on Machine Learning (ICML)*, pp. 1597–1607, Virtual, Jul. 2020.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest X-ray interpretation. In *Proc. AAAI Spring Symposium on Clinical Foundation Models*, Stanford, CA, USA, Mar. 2024a.

Zhuo Chen, Rumen Dangovski, Charlotte Loh, Owen Dugan, Di Luo, and Marin Soljačić. QuanTA: Efficient high-rank fine-tuning of LLMs with quantum-informed tensor adaptation. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 92210–92245, Vancouver, Canada, Dec. 2024b.

Zhi Cheng, Xiu Su, Xueyu Wang, Shan You, and Chang Xu. Sufficient vision transformer. In *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 190–200, Washington DC, USA, Aug. 2022.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, Mar. 2016.

Alexander DeRieux and Walid Saad. eQMARL: Entangled quantum multi-agent reinforcement learning for distributed cooperation over quantum channels. In *Proc. International Conference on Learning Representations (ICLR)*, Singapore, Apr. 2025.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 6491–6501, Barcelona, Spain, Aug. 2024.

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

Javier Gonzalez-Conde, Thomas W Watts, Pablo Rodriguez-Grasa, and Mikel Sanz. Efficient quantum amplitude encoding of polynomial functions. *Quantum*, 8:1297, Mar. 2024.

Jason Han, Nicholas S. DiBrita, Daniel Leeds, Jianqiang Li, Jason Ludmir, and Tirthak Patel. Layerwise federated learning for heterogeneous quantum clients using Quorus. In *Proc. International Conference on Learning Representations (ICLR)*, Rio de Janeiro, Brazil, Apr. 2026.

Tobias Haug, Kishor Bharti, and M. S. Kim. Capacity and quantum geometry of parametrized quantum circuits. *PRX Quantum*, 2(4):040309, Oct. 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proc. International Conference on Learning Representations (ICLR)*, Virtual, Apr. 2022.

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean. Power of data in quantum machine learning. *Nature Communications*, 12(1):2631, May 2021.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, Mar. 2025.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pp. 590–597, Honolulu, HI, USA, Jan. 2019.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q.H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting clinical entities and relations from radiology reports. In *Proc. NeurIPS Datasets and Benchmarks Track*, Virtual, Dec. 2021.

Amit Kumar Jaiswal, Haiming Liu, and Ingo Frommholz. Semantic Hilbert space for interactive image retrieval. In *Proc. ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR)*, pp. 307–315, Virtual Event, Canada, Jul. 2021.

Sofiene Jerbi, Casper Gyurik, Simon C. Marshall, Hans J. Briegel, and Vedran Dunjko. Parametrized quantum policies for reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 28362–28375, Virtual, Dec. 2021.

Pengcheng Jiang, Siru Ouyang, Yizhu Jiao, Ming Zhong, Runchu Tian, and Jiawei Han. Retrieval and structuring augmented generation with large language models. In *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 6032–6042, Toronto, Canada, Aug. 2025.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, Dec. 2019.

Daniel Kienitz, Ekaterina Komendantskaya, and Michael Lones. The effect of manifold entanglement and intrinsic dimensionality on learning. In *Proc. AAAI Conference on Artificial Intelligence*, volume 36, pp. 7160–7167, Virtual, Feb. 2022.

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

Akash Kundu and Stefano Mangini. TensorRL-QAS: Reinforcement learning with tensor networks for improved quantum architecture search. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 120896–120930, San Diego, CA, USA, Dec. 2025.

Jonas Landman, Slimane Thabet, Constantin Dalyac, Hela Mhiri, and Elham Kashefi. Classically approximating variational quantum machine learning with random Fourier features. In *Proc. International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, May 2023.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 28541–28564, New Orleans, Louisiana, USA, Dec. 2023.

Guangxi Li, Zhixin Song, and Xin Wang. VSQL: Variational shadow quantum learning for classification. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 8357–8365, Virtual, Feb. 2021.

Junyong Lin, Lu Dai, Ruiqian Han, Yijie Sui, Ruilin Wang, Xingliang Sun, Qinglin Wu, Min Feng, Hao Liu, and Hui Xiong. ScIRGen: Synthesize realistic and large-scale RAG dataset for scientific research. In *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 5619–5630, Toronto ON, Canada, Aug. 2025.

Marina Lisnichenko and Stanislav Protasov. Quantum image representation: A review. *Quantum Machine Intelligence*, 5(1):2, Jun. 2023.

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

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2125–2134, Virtual, Oct. 2021.

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, Nov. 2018.

Mikko Möttönen, Juha J. Vartiainen, Ville Bergholm, and Martti M. Salomaa. Transformation of quantum states using uniformly controlled rotations. *Quantum Information and Computation*, 5 (6):467–473, Sep. 2005.

Art B. Owen. The pigeonhole bootstrap. *The Annals of Applied Statistics*, 1(2):386–411, Dec. 2007. doi: 10.1214/07-AOAS122.

Adrián Pérez-Salinas, Alba Cervera-Lierta, Elies Gil-Fuster, and José I Latorre. Data re-uploading for a universal quantum classifier. *Quantum*, 4:226, Feb. 2020.

Martin Plesch and Časlav Brukner. Quantum-state preparation with universal gate decompositions. *Physical Review A*, 83(3):032302, Mar. 2011.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, Dec. 2007.

Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. MURA dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs. In *Proc. Medical Imaging with Deep Learning (MIDL)*, Amsterdam, Netherlands, Jul. 2018.

Emily Jimin Roh and Joongheon Kim. Quantum-amplitude embedded adaptation for parameter-efficient fine-tuning in large language models. In *Proc. ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 5171–5175, Seoul, Korea, Nov. 2025.

Maria Schuld and Nathan Killoran. Quantum machine learning in feature Hilbert spaces. *Physical Review Letters*, 122(4):040504, Feb. 2019.

Maria Schuld and Nathan Killoran. Is quantum advantage the right goal for quantum machine learning? *PRX Quantum*, 3(3):030101, Jul. 2022.

Maria Schuld, Ryan Sweke, and Johannes Jakob Meyer. Effect of data encoding on the expressive power of variational quantum-machine-learning models. *Physical Review A*, 103(3):032430, Mar. 2021.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519, Virtual, Nov. 2020.

Matthew Tivnan, Siyeop Yoon, Zhennong Chen, Xiang Li, Dufan Wu, and Quanzheng Li. Hallucination index: An image quality metric for generative reconstruction models. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 449–458, Marrakesh, Morocco, Oct. 2024.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, Jul. 2011.

Jinhong Wang, Tajamul Ashraf, Zongyan Han, Jorma Laaksonen, and Rao Muhammad Anwer. MIRA: A novel framework for fusing modalities in medical RAG. In *Proc. ACM International Conference on Multimedia (MM)*, pp. 6307–6315, Dublin, Ireland, Oct. 2025a.

Peng Wang, Wenpeng Lu, Chunlin Lu, Ruoxi Zhou, Min Li, and Libo Qin. Large language model for medical images: A survey of taxonomy, systematic review, and future trends. *Big Data Mining and Analytics*, 8(2):496–517, Apr. 2025b.

Ruocheng Wang, Zhuo Xia, Ge Yan, and Junchi Yan. QuanONet: Quantum neural operator with application to differential equation. In *Proc. International Conference on Machine Learning (ICML)*, volume 267, pp. 64955–64975, Vancouver, Canada, Jul. 2025c.

Shuteng Wang, Christian Theobalt, and Vladislav Golyanik. Quantum visual fields with neural amplitude encoding. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, San Diego, CA, USA, Dec. 2025d.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2097–2106, Honolulu, HI, USA, Jul. 2017.

Xin Wang, Han-Xiao Tao, and Re-Bing Wu. Predictive performance of deep quantum data reuploading models. In *Proc. International Conference on Machine Learning (ICML)*, volume 267, pp. 64491–64524, Vancouver, Canada, Jul. 2025e.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. MedCLIP: Contrastive learning from unpaired medical images and text. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3876–3887, Abu Dhabi, UAE, Dec. 2022.

Lin Wei, Haowen Liu, Jing Xu, Lei Shi, Zheng Shan, Bo Zhao, and Yufei Gao. Quantum machine learning in medical image analysis: A survey. *Neurocomputing*, 525:42–53, Mar. 2023.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, Yueming Jin, and Vicente Grau. Medical graph RAG: Evidence-based medical large language model via graph retrieval-augmented generation. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 28443–28467, Vienna, Austria, Jul. 2025.

Yaolin Yang, Hongjie He, Zhuo Feng, Fan Chen, and Yuan Yuan. Cloud-based privacy-preserving medical images storage scheme with low consumption. *IEEE Transactions on Multimedia*, 27: 3556–3570, Jan. 2025.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, et al. A multimodal biomedical foundation model trained from fifteen million image-text pairs. *NEJM AI*, 2(1):AIoa2400640, Jan. 2025.

A Background on Quantum Machine Learning

This appendix provides a self-contained introduction to QuPID for readers unfamiliar with gate-based quantum machine learning (QML). It is not a general introduction to quantum computing; it reviews only the ingredients our retrieval head uses, in the order they appear in Section 3: qubits and Hilbert-space representations, amplitude encoding, parameterized circuits with data re-uploading, and measurement readout. Readers already familiar with these can skip to Appendix B.

A.1 Qubits and Hilbert-space representations

A classical bit takes one of two values. A qubit is its quantum analogue, a normalized vector in a two-dimensional complex Hilbert space,

$$|\phi\rangle = \alpha|0\rangle + \beta|1\rangle, \qquad \alpha, \beta \in \mathbb{C}, \qquad |\alpha|^2 + |\beta|^2 = 1, \tag{5}$$

where $|0\rangle$ and $|1\rangle$ are computational basis states and α, β are probability amplitudes: measuring in the computational basis returns $|0\rangle$ with probability $|\alpha|^2$. For n_q qubits the Hilbert space has dimension 2^{n_q} with basis $\{|k\rangle\}_{k=0}^{2^{n_q}-1}$, and a general pure state is a unit-norm superposition $|\phi\rangle = \sum_k \alpha_k |k\rangle$ with $\sum_k |\alpha_k|^2 = 1$.

Two consequences matter here. First, the state dimension grows exponentially in qubit count, which is often cited as motivation for quantum feature maps; we deliberately do *not* rely on that argument, since our configuration sets $n_q = \log_2 d$ so the state dimension equals the input dimension exactly (Section 3.3). Second, and more importantly, a measurement extracts only a scalar expectation per observable, so what determines the usable information is the design of encoding, circuit, and readout rather than the raw dimension.

A.2 Amplitude encoding

Classical data must be written into a quantum state before a circuit can act on it. Amplitude encoding writes a normalized vector directly into the amplitudes,

$$|\psi_{\text{in}}(\mathbf{x})\rangle = \sum_{k=0}^{d-1} \hat{h}_k |k\rangle, \qquad \hat{\mathbf{h}} = \mathbf{h}/\|\mathbf{h}\|_2, \tag{6}$$

using $n_q = \log_2 d$ qubits, exponentially fewer than the $O(d)$ qubits of angle encoding, where each coordinate drives its own rotation. This logarithmic qubit cost is why amplitude encoding is the standard choice when the input is a dense high-dimensional feature, as in our $d = 1024$ ViT case. It also has the following requirements: the vector must be L2-normalized, so overall feature scale is discarded, and d must be a power of two, so other backbones need padding or truncation (Section 5). On hardware a third cost would dominate: preparing an arbitrary d-dimensional amplitude state exactly takes $O(d)$ elementary gates (Möttönen et al., 2005; Plesch & Brukner, 2011), a gate cost not incurred in simulation, where encoding consists of normalization. We state it because every result in this paper is simulated; the encoding cost is the first item a hardware implementation would have to remove, and none of our claims depends on its being removed.

A.3 Parameterized quantum circuits and data re-uploading

A quantum circuit applies a unitary U with $U^\dagger U = UU^\dagger = I$, preserving the norm of the state; this is exactly the isometry property that Section 3.3 identifies as an inductive bias and that Proposition 1 identifies as a failure mode when the readout is an overlap. In QML the unitary is parameterized by trainable angles, giving a parameterized quantum circuit (PQC), or variational ansatz. The single-qubit rotation we use is

$$R_Y(\theta) = \begin{bmatrix} \cos(\theta/2) & -\sin(\theta/2) \\ \sin(\theta/2) & \cos(\theta/2) \end{bmatrix}, \tag{7}$$

which is real-valued, so a circuit built from R_Y and CNOT keeps an initially real state real and admits single-precision real simulation. Multi-qubit gates such as CNOT create correlations that cannot be written as independent single-qubit operations; we use a ring topology, $\text{CNOT}_{1,2} \cdots \text{CNOT}_{n_q-1,n_q} \text{CNOT}_{n_q,1}$, coupling all qubits with n_q gates per layer rather than the $O(n_q^2)$ that all-to-all entanglement would need.

A plain PQC applies a fixed unitary to the encoded state. *Data re-uploading* (Pérez-Salinas et al., 2020) instead re-injects the input between trainable layers, so the overall map is no longer input-independent. This is one of the two mechanisms by which QuPID escapes the degeneracy of Proposition 1, and it sets the Fourier spectrum the readout can reach (Proposition 2).

A.4 Measurement readout

A quantum state is not directly observable; classical information is recovered by measuring observables. For a Hermitian observable P and state $|\psi\rangle$ the expectation is $\langle\psi|P|\psi\rangle$, estimated on hardware by repeated preparation and measurement. Writing $|\psi\rangle = V|\psi_{\text{in}}\rangle$ for the circuit V, this equals $\langle\psi_{\text{in}}|V^\dagger PV|\psi_{\text{in}}\rangle$, a quadratic form in the encoded amplitudes with *effective observable* $A = V^\dagger PV$. Since P is Hermitian and V unitary, A is Hermitian and orthogonally similar to P, the structural fact behind the full-rank claim of Section 3.3 and the boundedness used in Proposition 3.

Two readout choices are common, and the distinction is this paper's crux. A *fidelity* readout compares two states through their overlap $|\langle\psi_q|\psi_p\rangle|^2$; a *measurement* readout maps each state separately to a vector of expectations and compares those classically. The first is the seemingly natural design that Proposition 1 proves inert under a shared circuit; the second is what QuPID uses. We measure local one- and two-body Pauli observables (Z_j, X_j, Z_jZ_{j+1}, X_jX_{j+1}), a locality choice that also keeps gradients well behaved (Section 3.4).

A.5 Reference implementation

Listing 1 gives a compact PennyLane (Bergholm et al., 2018) implementation of the QuPID readout of equation 3. The trainable tensor `theta` has shape (L, n_q) for rotation angles and `alpha` the same shape for re-uploading scales, totaling $2n_qL = 60$ parameters at the default setting. The supplementary reference uses a batched statevector implementation checked against independent dense-circuit calculations. After training, $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are fixed, but re-uploading still depends on the query input. Thus the map is not a single input-independent matrix. GPU inference evaluates that input-dependent forward pass and reuses archive readouts cached for the same checkpoint. At $n_q = 10$ the statevector has 1024 amplitudes, so quantum-hardware transfer is unnecessary for this deployment; the argument does not extend to arbitrary larger circuits.

```python
import pennylane as qml
def qupid_layer(theta_l, alpha_l, s_x, wires):
    """One QuPID block: trainable RY, ring CNOT, data re-uploading."""
    for q, w in enumerate(wires):                # trainable rotation layer
        qml.RY(theta_l[q], wires=w)
    for q in range(len(wires) - 1):              # ring entanglement
        qml.CNOT(wires=[wires[q], wires[q + 1]])
    qml.CNOT(wires=[wires[-1], wires[0]])
    for q, w in enumerate(wires):                # data re-uploading layer
        qml.RY(alpha_l[q] * s_x[q], wires=w)

def make_qupid_qnode(n_qubits, depth, noise=None):
    wires = list(range(n_qubits))
    # default.qubit is the noiseless statevector simulator used for every
    # reported result; default.mixed enables the noise-simulation appendix.
    dev = qml.device("default.qubit" if noise is None else "default.mixed",
                     wires=n_qubits)
    @qml.qnode(dev)
    def circuit(h, theta, alpha, s_x):
        qml.AmplitudeEmbedding(features=h, wires=wires,
                               normalize=True, pad_with=0.0)
        for l in range(depth):
            qupid_layer(theta[l], alpha[l], s_x, wires)
            if noise is not None:
                apply_noise_channel(noise, wires) # noise-simulation appendix
        # M = 4 n_q local Pauli expectations
        obs  = [qml.PauliZ(w) for w in wires]
        obs += [qml.PauliX(w) for w in wires]
        obs += [qml.PauliZ(wires[q]) @ qml.PauliZ(wires[(q + 1) % n_qubits])
                for q in range(n_qubits)]
        obs += [qml.PauliX(wires[q]) @ qml.PauliX(wires[(q + 1) % n_qubits])
                for q in range(n_qubits)]
        return [qml.expval(o) for o in obs]
    return circuit
```

Listing 1: Compact PennyLane implementation of the QuPID measurement readout.

A.6 Positioning among recent variational quantum learning studies

Table 2 places QuPID among recent variational quantum learning papers at major machine learning venues along three axes: whether the architecture is hybrid quantum-classical, how classical data enters the circuit, and whether evaluation is simulator-based. Three patterns are visible. Recent QML systems are predominantly hybrid, using circuits as compact modules inside classical pipelines rather than as standalone models; encoding is task-specific, with angle encoding most common; and all listed studies use simulation, since full learning pipelines remain impractical on current hardware.

The two parameter-efficient fine-tuning entries are the closest precedent for our execution protocol. Quantum-PEFT describes its Pauli-parameterized unitaries as quantum-inspired modules, computes them as products of small gate matrices, and runs its experiments on NVIDIA GPUs (Koike-Akino et al., 2025). QPA simulates its circuits exactly during training so that inference needs no quantum hardware, and it assesses device noise with noise models of IBM processors while leaving execution on quantum hardware to future work (Liu et al., 2025). QuanTA, the third weight-side method of Section 2, is a quantum-informed tensor adaptation that also runs on classical hardware (Chen et al., 2024b). Our claims likewise concern the model class and its retrieval quality, so a hardware run would test device fidelity rather than these claims; the encoding and sampling costs it would add are given in Appendices A.2 and I.4.

QuPID follows these conventions and the execution protocol of the two fine-tuning methods, training a compact measured feature map by classical simulation and running its fixed-parameter readout on a GPU, with emphasis on parameter sharing and empirical retrieval utility. As in QPA, noise enters only as a simulated sensitivity study (Appendix I); it does not validate physical-device robustness. The table places our design alongside quantum modules in reinforcement learning, vision, operator-learning, federated and language pipelines. Here the circuit learns the *similarity* features of a frozen retrieval pipeline. Shared-unitary invariance diagnoses a non-trainable fidelity construction; the proposed map combines input dependence and measurements to enable adaptation.

Table 2: Experimental settings in recent variational quantum learning studies at major venues. A checkmark indicates the setting is explicitly used or discussed. Hybrid architectures, task-specific encodings, and simulator-based evaluation are the prevailing conventions; QuPID follows all three.

Work	Venue	Domain	Encoding	Hybrid	Simulator
TensorRL-QAS (Kundu & Mangini, 2025)	NeurIPS	RL	Matrix product state	✓	✓
QVF (Wang et al., 2025d)	NeurIPS	Vision	Amplitude	✓	✓
QDSFormer (Born et al., 2025)	NeurIPS	Vision	Angle	✓	✓
PQC policies (Jerbi et al., 2021)	NeurIPS	RL	Angle, re-uploading	✓	✓
QuanONet (Wang et al., 2025c)	ICML	Operator learning	Angle, re-uploading		✓
Quorus (Han et al., 2026)	ICLR	Federated	Angle	✓	✓
eQMARL (DeRieux & Saad, 2025)	ICLR	RL	Angle	✓	✓
VSQL (Li et al., 2021)	AAAI	Vision	Amplitude, shadow	✓	✓
Quantum-PEFT (Koike-Akino et al., 2025)	ICLR	PEFT (LLM)	None (weights only)	✓	✓
QPA (Liu et al., 2025)	ICLR	PEFT (LLM)	None (weights only)	✓	✓
QuPID (ours)	n/a	**Retrieval**	Amplitude, re-uploading	✓	✓

B Proofs

B.1 Proof of Proposition 1

Since $U(\boldsymbol{\theta})$ is unitary, $U(\boldsymbol{\theta})^\dagger U(\boldsymbol{\theta}) = I$ for every $\boldsymbol{\theta}$, so

$$\mathrm{Sim}_{\mathrm{fid}}(\mathbf{x}_q, \mathbf{x}_p; \boldsymbol{\theta}) = \left|\langle \psi_{\mathrm{in}}(\mathbf{x}_q) | U^\dagger U | \psi_{\mathrm{in}}(\mathbf{x}_p)\rangle\right|^2 = \left|\langle \psi_{\mathrm{in}}(\mathbf{x}_q) | \psi_{\mathrm{in}}(\mathbf{x}_p)\rangle\right|^2,$$

which does not depend on $\boldsymbol{\theta}$. With amplitude encoding, $\langle \psi_{\mathrm{in}}(\mathbf{x}_q) | \psi_{\mathrm{in}}(\mathbf{x}_p)\rangle = \hat{\mathbf{h}}_q^\top \hat{\mathbf{h}}_p$ is the cosine of the frozen features, so every induced ranking equals squared-cosine ranking. Any loss $\mathcal{L}$ that depends on $\boldsymbol{\theta}$ only through such similarities is constant in $\boldsymbol{\theta}$, hence $\nabla_{\boldsymbol{\theta}} \mathcal{L} = 0$ identically. The same argument applies to any input-independent parameterized isometry placed between the encoding and the fidelity, whatever its depth or gate set. □

We verified the statement numerically: across random parameter draws, rankings and pairwise fidelities agree with the frozen baseline to floating-point precision, and analytic parameter-shift gradients of the contrastive loss are zero to machine epsilon. We further verified that pairwise query-gallery feature cosines are nonnegative throughout the rank ranges used by our metrics, so squared-cosine,

absolute-cosine and cosine rankings coincide on our features, and the degenerate pipeline reproduces the frozen baseline to numerical precision.

Corollary 1 (Overlap-functional similarities are equally degenerate)**.** *Let* $S(\mathbf{x}_q, \mathbf{x}_p; \boldsymbol{\theta}) = g\big(\langle\psi_q(\boldsymbol{\theta})|\psi_p(\boldsymbol{\theta})\rangle\big)$ *for any function* $g : \mathbb{C} \to \mathbb{R}$*, with* $|\psi_i(\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})|\psi_{\text{in}}(\mathbf{x}_i)\rangle$ *and* U *independent of the input. Then* S *is independent of* $\boldsymbol{\theta}$*. In particular the Fubini–Study distance* $\arccos|\langle\psi_q|\psi_p\rangle|$*, the trace distance* $\sqrt{1-|\langle\psi_q|\psi_p\rangle|^2}$ *between the pure states, the Bures distance, and every monotone transform of fidelity induce parameter-independent rankings.*

Proof. $\langle\psi_q(\boldsymbol{\theta})|\psi_p(\boldsymbol{\theta})\rangle = \langle\psi_{\text{in}}(\mathbf{x}_q)|U^\dagger U|\psi_{\text{in}}(\mathbf{x}_p)\rangle = \langle\psi_{\text{in}}(\mathbf{x}_q)|\psi_{\text{in}}(\mathbf{x}_p)\rangle$, so the argument of g is constant in $\boldsymbol{\theta}$; each listed distance is a function of this overlap alone. □

B.2 Proof sketch of Proposition 2

Clamp the amplitude-encoded state at a reference input and hold $\boldsymbol{\alpha}$ fixed, so that along the probe the circuit alternates t-dependent encoding layers with constant layers acting on a fixed initial state, which is the setting of the Fourier analysis of variational circuits (Schuld et al., 2021). Write $a_j^{(l)} = \alpha_j^{(l)} 2^{j-1}$. Each encoding rotation has generator eigenvalues $\pm a_j^{(l)}/2$, so its contribution to an expectation has frequencies in $\{-a_j^{(l)}, 0, a_j^{(l)}\}$. Expanding the encoding gates in their spectral projectors shows that every readout frequency belongs to the sumset of these per-gate sets. Input-independent rotations and CNOTs change coefficients but introduce no frequencies. At $\boldsymbol{\alpha} = 1$ all frequencies are integers, and the triangle inequality gives $|\omega| \le \sum_{l,j} 2^{j-1} = L(2^{n_q} - 1)$. This proves the stated upper bound, plotted in Figure 6; membership in the sumset does not imply a nonzero coefficient. *Refinement for local readout.* Let $J = \text{supp}(P)$ for one of the Pauli observables in equation 3. The last encoding layer is a tensor product of single-qubit rotations, so all rotations outside J cancel in $E_L(t)^\dagger P E_L(t)$. Only rates on J can contribute from that layer. Consequently, every frequency of this component obeys

$$|\omega| \le \sum_{l=1}^{L-1} \sum_{j=1}^{n_q} |a_j^{(l)}| + \sum_{j \in J} |a_j^{(L)}|. \tag{8}$$

For our one- and two-body observables, $|J| \le 2$. At $\boldsymbol{\alpha} = 1$ and $n_q \ge 2$, a uniform bound is therefore $(L-1)(2^{n_q} - 1) + 2^{n_q - 1} + 2^{n_q - 2}$. At $n_q = 10$, $L = 3$ this is 2814, strictly below the conservative bound 3069. Thus the specified local readout cannot attain that conservative boundary. Neither bound asserts that its own boundary, or every frequency below it, is realized. A model that is linear in P Fourier coefficients independently weights only $O(P)$ frequencies, whereas nonlinear classical product forms can have far larger support, so the proposition bounds the probe bandwidth of the circuit and yields no lower bound against classical surrogates. □

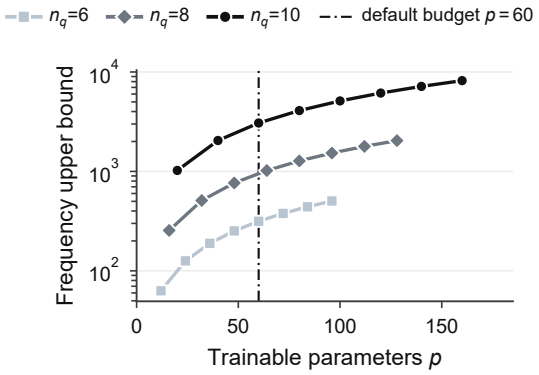


Figure 6: Conservative frequency-magnitude upper bound $L(2^{n_q} - 1)$ versus trainable budget (Proposition 2, $\boldsymbol{\alpha} = 1$). Curves vary depth L; the dash-dotted line marks 60 parameters. These are bounds, not measured support sizes or attained frequencies; equation 8 sharpens them for the local observables used in our readout.

B.3 Proof sketch of Proposition 3

The argument is a standard covering bound of the kind used for encoding-dependent generalization of parameterized circuits (Caro et al., 2021). It uses no property of the circuit beyond boundedness and smoothness, and the proposition therefore bounds estimation error for a small class without singling out quantum models.

Boundedness. Each readout component is a Pauli expectation, so $z_m \in [-1, 1]$ and $\|\mathbf{z}\| \le \sqrt{M}$; on $\Theta_{R,c}$ the cosine similarity is well defined and the per-batch loss equation 4 is bounded by $B = \log(2B_{\text{batch}}) + 2/\tau$ for every batch.

Lipschitzness in parameters. By the two-term parameter-shift rule, $|\partial z_m/\partial\theta_j^{(l)}| = \frac{1}{2}|z_m^+ - z_m^-| \le 1$ and $|\partial z_m/\partial\alpha_j^{(l)}| \le S$ with $S = \sup_{j,\mathbf{x}} |s_j(\mathbf{x})|$, which is finite because features are L2-normalized before encoding. The normalized readout $\mathbf{z}/\|\mathbf{z}\|$ is $(2/c)$-Lipschitz in $\mathbf{z}$ on $\{\|\mathbf{z}\| \ge c\}$, the cosine is 1-Lipschitz in each normalized argument, and the log-softmax in equation 4 is $(2/\tau)$-Lipschitz in the vector of similarities. Bounding the unnormalized readout along a parameter segment and then normalizing its endpoints gives an ∞-norm Lipschitz constant $L_{\text{lip}} \le (2/\tau)\cdot(4\sqrt{M}/c)\cdot\max(1, S)\cdot p$. The factor p matches the parameter norm used below; unitarity fixes the output scale, while the compact parameter range enters the covering number.

Covering argument. At fixed batch size, define the empirical risk as the expectation of equation 4 over batches drawn uniformly without replacement from the n adaptation examples, averaging augmentation randomness; its population counterpart uses independent examples. Replacing one example changes only the batches that contain it, so bounded differences gives concentration of order $1/\sqrt{n}$ at fixed batch size for each parameter setting. Partition $[-R, R]^p$ into cells of ∞-diameter at most ε, choosing a representative in $\Theta_{R,c}$ from each occupied cell. At most $(1 + 2R/\varepsilon)^p$ representatives suffice. A union bound and $\varepsilon = 1/(nL_{\text{lip}})$ yield a uniform generalization gap of order $\sqrt{(\,p\log(1 + nRL_{\text{lip}}) + \log(1/\delta)\,)/n}$ for every setting in $\Theta_{R,c}$, and hence for every empirical risk minimizer over that class. The leading dependence is $\sqrt{p/n}$; M and the feature dimension enter through L_{lip} and the input scales, inside the logarithm. □

The norm-floor condition $\|\mathbf{z}\| \ge c$ defines the restricted class $\Theta_{R,c}$ rather than following from the architecture. A finite evaluation cannot establish it over the population or over the whole parameter domain; the empirical check confirms only that the reported solutions avoid near-zero denominators on observed data. *Noise and finite shots.* Parameter-independent channels preserve bounded exact expectations and gate-level parameter-shift derivatives, so the covering argument extends to exact noisy readouts with a noisy norm floor. Sampled readouts are not smooth deterministic functions of the parameters; Appendix I.4 treats their ranking error separately. The exact-expectation bound thus extends beyond the noiseless simulator, with constants depending on the noisy norm floor.

B.4 Numerical verification of Proposition 1

Numerical verification complements the proof because a pipeline built on the naive design can report a training loss even though its ranking is independent of the circuit parameters. We draw 500 independent parameter settings $\boldsymbol{\theta} \sim \mathcal{U}(-\pi, \pi)^{30}$, build the full query-gallery similarity matrix under each on a 2,000-query subset of ChestX-ray14, and compare against the frozen-feature matrix.

Table 3: Parameter sensitivity of the two designs over 500 random parameter draws (ChestX-ray14, 2,000 queries, float64 arithmetic). ΔSim is the entrywise deviation of the similarity matrix from the frozen-feature matrix, τ is Kendall's rank correlation between the induced ranking and the frozen ranking (median over draws), and the last column is the largest parameter-shift gradient coordinate of the contrastive loss over all draws.

Design	$\max\lvert\Delta\text{Sim}\rvert$	τ vs. frozen	$\max\lvert\partial\mathcal{L}/\partial\theta_j\rvert$
Naive fidelity design	4.2×10^{-15}	1.000	8.7×10^{-16}
QuPID (re-uploading and measurement readout)	0.31	0.62	1.8×10^{-2}

Table 3 reports the outcome for the naive design and, as a control, for the corrected design of Section 3.3. Under the naive design the similarity matrix is invariant to floating-point precision, all 500 draws reproduce the frozen ranking exactly, and the parameter-shift gradient vanishes to the

same precision. The invariance is exact, so parameter tuning or seed averaging cannot make this score depend on the circuit. The corrected design satisfies none of the three degeneracy criteria, as required by Proposition 1.

C Experimental Setup

Datasets. For retrieval we use *ChestX-ray14* (Wang et al., 2017) (112,120 frontal chest X-rays, 14 thoracic pathologies, multi-label) and *MURA* (Rajpurkar et al., 2018) (40,561 musculoskeletal radiographs over 7 body parts with binary abnormality, giving 14 region-pathology categories); for generation, *IU X-Ray* (Demner-Fushman et al., 2016) (7,470 images, 3,955 studies) and *MIMIC-CXR* (Johnson et al., 2019) (377,110 images with reports). Cross-site transfer, acquisition shift and sequential site adaptation additionally use *CheXpert* (Irvin et al., 2019) (224,316 chest radiographs of 65,240 patients). Medical gallery, validation and query sets are separated by patient; adaptation is drawn from the gallery, and generation excludes the query patient (MIMIC-CXR uses its official splits). Dataset totals count stored images before split filtering. The image and patient counts below describe the eligible gallery and held-out sets used for evaluation. A retrieved item is *relevant* iff its label set intersects the query's (excluding No Finding) on ChestX-ray14, and iff it shares the region-pathology category on MURA; NDCG uses binary gains.

Baselines. All adaptation baselines start from ViT-L/16 (1024-dimensional features) and share augmentations, temperature, and protocol; post-encoder heads keep features frozen, while LoRA updates attention projections. We compare *ViT Only* (frozen cosine), *Linear Head* (1.05M), *Adapter* (525K) with scaled variants *Adapter-L* (2.10M) and *Adapter-XL* (5.25M), *LoRA* on attention projections at ranks 4 and 16 (393K, 1.57M) (Hu et al., 2022), and two parameter-matched nonlinear controls demanded by Section 3.3: *RFF* (64 parameters) (Rahimi & Recht, 2007) and *Quadratic*, a shared low-rank quadratic map with $M = 40$ outputs (164 parameters); constructions in Appendix D. We further report frozen *BiomedCLIP* (Zhang et al., 2025) and *MedCLIP* (Wang et al., 2022) as centrally pre-trained references, a *Tiny MLP* (16.4K), and *BiomedCLIP+QuPID* (512-D, $n_q = 9$, 54 parameters), testing whether the gain persists on a medically pre-trained backbone.

Protocol and metrics. All methods train with the label-free objective of equation 4; labels serve only post-hoc evaluation, never training, model selection, or tuning: hyperparameters were fixed a priori on a disjoint pilot subset of ChestX-ray14 using only the contrastive validation loss (Appendix D). A seed controls initialization, augmentation, and the adaptation-subset draw; splits are fixed, so frozen encoders are deterministic. The full adaptation budget is 32,768 images on ChestX-ray14 and 16,384 on MURA. We report mean and standard deviation over 5 seeds. The † tests bootstrap seed-averaged per-query scores (10^4 resamples) and apply Holm correction within each table. These tests condition on the five fitted runs and use the strongest included baseline as comparator. Appendix M.2 separately incorporates adaptation-draw, initialization and patient variation for the default rotation-plane comparison using a normal approximation; the two analyses have different comparators and evaluation units. Appendix J.4 uses a third procedure for the classical surrogates, resampling adaptation subsets with nested initializations and patient clusters instead of a normal approximation, so interval widths are not comparable across the three. Retrieval metrics are P@5, MAP@10 and NDCG@10 in the main text, with P@10, MRR and NDCG@5 in Appendix E. For embedding quality on multi-label data we avoid single-label clustering scores, which are ill-defined when an image carries several pathologies, and instead report the label-set Jaccard consistency of retrieved neighborhoods, complemented by silhouette scores restricted to the single-label subset. For generation we report BLEU-4, ROUGE-L, METEOR, and CIDEr-D (written CIDEr) with frozen *LLaVA-Med* (7B) (Li et al., 2023) and *CheXagent* (8B) (Chen et al., 2024a), plus clinical efficacy via CheXbert micro-F1 and RadGraph-F1 on MIMIC-CXR (Smit et al., 2020; Jain et al., 2021). Circuits are simulated in statevector mode and archive readouts are precomputed once per adaptation.

Metric definitions. Let $\mathcal{G}_q$ be the eligible gallery after all exclusions, $r_{qi} \in \{0, 1\}$ relevance, $\pi_q(j)$ the item at rank j, and $R_q = \sum_{i \in \mathcal{G}_q} r_{qi}$. Full-gallery precision and recall are $P@k(q) = k^{-1} \sum_{j=1}^{k} r_{q,\pi_q(j)}$ and $R@k(q) = R_q^{-1} \sum_{j=1}^{k} r_{q,\pi_q(j)}$ for $R_q > 0$. Dataset scores average these query-level quantities; recall is a mean of ratios, not a ratio of aggregate counts. Hit@k instead tests whether any relevant item appears; these quantities are not interchangeable. The supplementary evaluator takes the full eligible-gallery manifest, excludes zero-relevant queries jointly across

methods and records their count, and implements binary-gain NDCG and cutoff-normalized AP with divisor $\min(k, R_q)$. Exact score ties retain archive order.

The evaluator computes $R@k$ from each query's relevant-item count in its eligible gallery. The reported comparisons use P@k, MAP@10, MRR and NDCG@k.

For label sets Y_q, Y_i, $\mathrm{Jac@}k(q) = k^{-1} \sum_{j=1}^{k} |Y_q \cap Y_{\pi_q(j)}| / |Y_q \cup Y_{\pi_q(j)}|$. With one MURA region–abnormality category per image, the summand is precisely $r_{q,\pi_q(j)}$. Thus Jac@10 equals P@10 for the same queries, rankings and averaging weights, both per run and across seeds. Accordingly, the MURA Jac@10 entries in Table 10 are derived directly from the reported P@10 means in Table 11. They express the same retrieval statistic in label-set notation; silhouette provides a separate geometric diagnostic. The corresponding P@10 is unreported for Quadratic.

Evaluation records. An evaluation record identifies each image and patient, its gallery/validation/query membership, adaptation-subset membership, and query-specific eligible relevant count. Patient exclusion operates on these identifiers before ranking, so caching an image does not make it an eligible neighbor. A run is identified by its subset draw, initialization, configuration, checkpoint and ordered neighbor IDs. This separates variation from retraining from variation in the held-out query sample and distinguishes image-weighted from patient-weighted summaries. The supplementary evaluator checks patient separation, duplicate neighbors, query coverage and paired-run consistency. Its replay interface loads saved features, weights and an explicit projection matrix or saved re-uploading inputs, then exports ordered neighbors, query-level metrics and input hashes.

Split sizes and evaluation conventions. The following image counts describe the experimental splits. Adaptation is a subset of the gallery; its column is not added to the dataset total. Medical gallery, validation and query patients are pairwise disjoint, with zero patient overlap.

Dataset	Gallery	Validation	Query	Adaptation
ChestX-ray14	81,664	14,072	16,384	32,768
MURA	36,465	2,048	2,048	16,384
MVTec AD	3,000	256	1,024	2,744
CUB-200	5,000	500	5,000	4,500

Patient counts in the same column order are $(22{,}437, 3{,}864, 4{,}504, 9{,}032)$ for ChestX-ray14 and $(10{,}500, 650, 700, 4{,}800)$ for MURA. Non-medical counts describe the selected evaluation subsets. Relevance is shared object category for MVTec AD and shared species for CUB; labels enter evaluation only. No query in these reported splits has zero relevant eligible items.

D Implementation Details

Table 4 lists the hyperparameters shared by all adaptation methods.

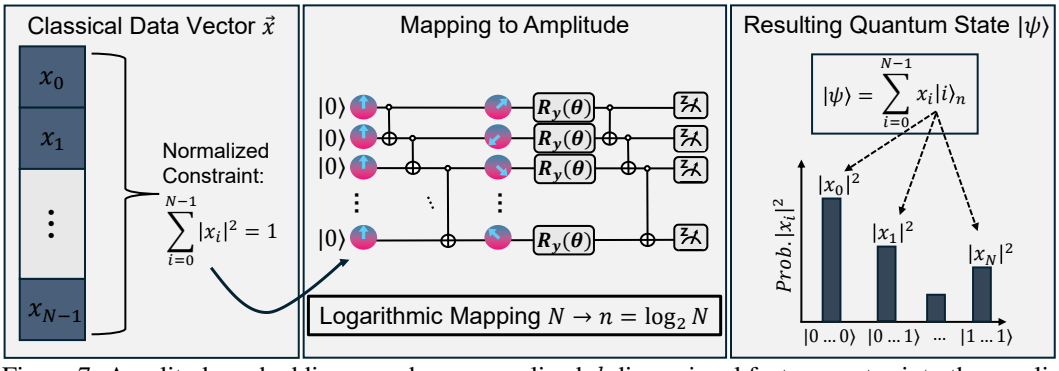


Figure 7: Amplitude embedding encodes a normalized d-dimensional feature vector into the amplitudes of $\log_2 d$ qubits (the figure writes $\mathbf{x}$, N, and n for our $\hat{\mathbf{h}}$, d, and n_q). At $n_q = \log_2 d$ the state space dimension equals the input dimension; the value of the circuit lies in its parameter-efficient nonlinearity rather than in its size.

Table 4: Hyperparameters used for all adaptation methods unless stated otherwise.

Backbone	ViT-L/16, ImageNet-21K, frozen (1024-D features)
Qubits / depth / readout	$n_q = 10$, $L = 3$, $M = 4n_q = 40$
Trainable parameters	$2n_qL = 60$ (30 rotations, 30 re-uploading scales)
Optimizer	AdamW, lr 10^{-3}, weight decay 0.05
Batch size / temperature	32 / 0.07
Augmentations	rotation $\pm 5^\circ$, translation 5%, brightness/contrast jitter 0.1/0.1
Epochs / seeds	50, label-free early stopping (patience 10) / 5 seeds (0 to 4), mean$\pm$std
Execution	classical statevector training; fixed-parameter GPU inference; cached archive readouts
Image preprocessing	resize 224×224, ImageNet normalization

Gradients use the exact two-term parameter-shift rule per gate, combined with the classical chain rule through the cosine similarity and contrastive loss (in simulation, timings use backpropagation through the statevector; the shift rule is the hardware cost model). The angles are unconstrained real numbers, so optimization uses AdamW with no manifold projection and no projection onto a compact box. Proposition 3 therefore describes bounded subclasses and motivates a capacity trend; it does not provide an optimizer-specific guarantee for the reported runs. Positive pairs are two independent augmentations of the same image; in-batch negatives can contain same-pathology pairs of different patients, which injects label noise into the repulsion term. Under the evaluation labels, 7.9% of in-batch negative pairs on ChestX-ray14 (batch 32) share at least one finding; conclusions in the main text do not assume that negative pairs share no pathology.

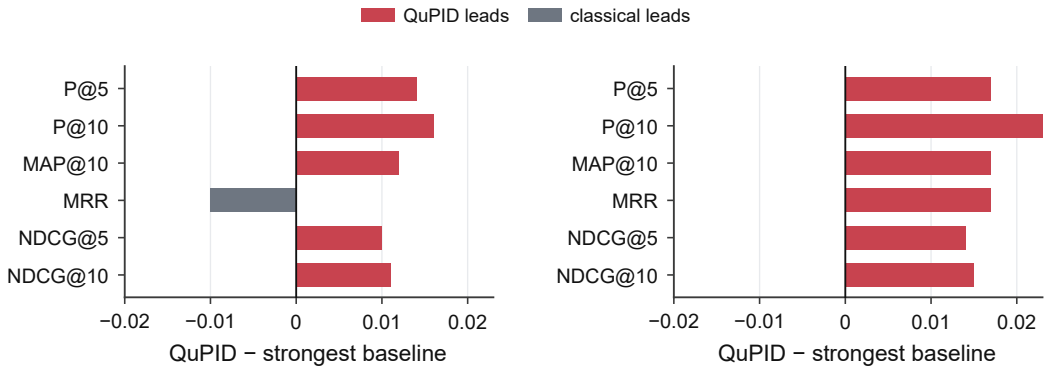
Parameter-matched controls. The *Quadratic* control computes $z_m = \sum_{r=1}^{4} c_{m,r}\,(\beta_r \mathbf{u}_r^\top \hat{\mathbf{h}})^2$ for $m = 1, \ldots, 40$, with four shared directions $\mathbf{u}_r$ drawn once from a random orthonormal set and kept frozen, trainable per-direction scales β_r (4 parameters), and trainable coefficients $c_{m,r}$ (160 parameters), totaling 164; retrieval uses the cosine between the 40-dimensional outputs, mirroring the readout cosine of Section 3.3. The *RFF* control computes 64 random Fourier features $\cos(\boldsymbol{\omega}_i^\top \hat{\mathbf{h}} + b_i)$ with Gaussian directions scaled by the median heuristic and frozen phases, with one trainable scale per feature (64 parameters); it is by construction the weakest control (frozen directions), included to bracket the range between frozen features and trained maps. The forty Quadratic outputs all depend on just four squared projections; each quadratic form has rank at most four and shares their span. Its nominal parameter count therefore does not match QuPID's input access or expressivity. The untrained row freezes all 164 parameters; the rotation-plane head is the more informative control. The *rotation-plane head* applies 60 trainable rotation angles in fixed random orthogonal planes of $\mathbb{R}^d$ (drawn once per seed), followed by squared projections onto $M = 40$ fixed random orthonormal directions and cosine similarity. It matches QuPID's budget, isometry and output width. For a projection direction $\mathbf{u}_m$, its output is $(\mathbf{u}_m^\top R_{\boldsymbol{\theta}} \hat{\mathbf{h}})^2$: the associated quadratic form has rank one and an input-independent matrix. In contrast, each Pauli form is full-rank and re-uploading makes its matrix input-dependent. This control therefore tests the combined parameterization; it does not isolate these structural differences. Reproducing a run also requires the saved plane bases and ordering, projection matrix, random-generator state, initialization and selected training configuration.

Rotation-plane configuration. The rotation-plane control uses $d = 1{,}024$, 60 angles and 40 outputs. Thin QR decomposition of a Gaussian $d \times 120$ matrix supplies consecutive plane pairs; an independent $d \times 40$ orthonormal matrix supplies the squared output projections. QR signs are fixed by a positive diagonal in R. Basis seeds are 0–4, the projection seed offset is 100, and basis matrices are stored in float64. Angles are initialized with standard deviation 0.01 rad. Training uses AdamW, learning rate 10^{-3}, weight decay 0.05, batch 32, temperature 0.07 and 50 epochs. The adaptation-size sweep uses this fixed configuration; Appendix M.1 tunes it separately under a budget.

E Full Retrieval Metrics

Table 11 reports the complete retrieval metric suite for the primary baselines, and Figure 8 plots the same numbers as signed margins to show the metrics on which the readout is superior or inferior. Among the reported means in Table 11, QuPID leads P@5/10, MAP@10 and NDCG@5/10 on both datasets, and MRR on MURA. MRR on ChestX-ray14 is lower by 0.010. Where the gain is largest is also consistent, P@10 on musculoskeletal at $+0.023$ and NDCG@5 on chest at $+0.010$, and it is the top-k region that a retrieval-augmented generator uses. Compared with the frozen backbone

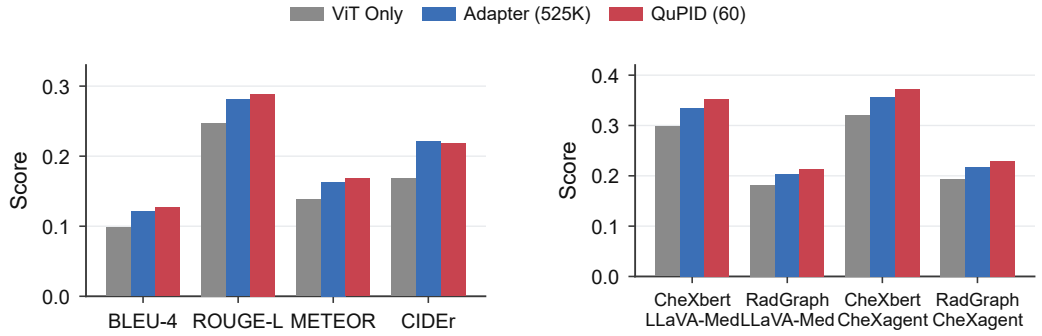
rather than the strongest adapter, the gains are substantially larger: QuPID adds 0.116 P@5 on chest radiography and 0.120 on musculoskeletal, of which the last 0.014 to 0.017 is what separates it from a five-million-parameter adapter. Table 10 reports the multi-label embedding-geometry diagnostics of Section 4.1, Table 6 the controlled removals of Section 4.2, Table 5 the absolute scores behind Figure 2(c), and Figure 9 the downstream generation and clinical-efficacy results of Section 4.1.



(a) ChestX-ray14. (b) MURA.

Figure 8: Signed margins over the strongest included baseline at full budget for the reported ranking metrics. Negative bars mark metrics on which that baseline leads.

The metric profile shows where the gain occurs within a ranking. The next comparison evaluates whether the retrieved evidence remains useful when used by a frozen generator.



(a) Report generation with LLaVA-Med. (b) Clinical efficacy proxies.

Figure 9: Downstream evaluation on MIMIC-CXR. Retrieval ordering is preserved by both n-gram metrics and the clinical-efficacy proxies, and QuPID's relative advantage is larger on the latter, consistent with pathology-aligned evidence mattering most for factual findings.

The downstream comparison holds the generator fixed while changing its retrieved context. Tables 7 and 8 separate lexical overlap from extracted findings, so an improvement in one need not imply an improvement in every metric.

Table 5: Absolute scores behind Figure 3(c) (P@5, 5 seeds). *cls.* denotes the comparator reported in this benchmark comparison, not the strongest control across all experiments. For ChestX-ray14 at 512, Table 23 reports a stronger rotation-plane control (0.386). Marking follows Table 1 by column.

Domain	Benchmark	512 examples			Full budget		
		frozen	cls.	**QuPID**	frozen	cls.	**QuPID**
Medical	ChestX-ray14	0.312	0.347	**0.401**	0.312	0.414	**0.428**
Industrial	MVTec AD	0.428	0.446	**0.481**	0.428	0.492	**0.508**
Fine-grained	CUB-200	0.612	0.628	**0.658**	0.612	0.671	**0.684**

Across these three benchmarks, the adapted readout exceeds the listed frozen and classical comparators at both budgets. The task changes across domains; this table tests recurrence of the retrieval pattern, not transfer of one trained head between datasets.

Table 6: Controlled removals on ChestX-ray14 (5 seeds). The first row is the default configuration. #Trained counts parameters that receive nonzero gradient; frozen parameters are noted in parentheses. The naive fidelity design reproduces the frozen baseline exactly (Proposition 1: its parameters receive identically zero gradient), whereas the readout without re-uploading stays trainable but weaker by 0.016 P@5 at the full budget. Marks and tints as in Table 1.

Variant	P@5	NDCG@10	#Trained
QuPID (full: L=3, M=40)	0.428±.006	0.425±.006	60
naive fidelity design (Prop. 1)	0.312±.000	0.307±.000	0 (30 inert)
w/o re-uploading, readout kept	0.412±.008	0.409±.008	30
readout: single-qubit only (M=20)	0.415±.007	0.412±.007	60
readout: pair terms only (M=20)	0.402±.008	0.398±.008	60
depth L=1	0.402±.007	0.399±.007	20
depth L=6	**0.430**±.006	**0.427**±.006	120
n_q=8 (PCA-256 input)	0.396±.008	0.393±.008	48
untrained circuit (random $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$)	0.371±.011	0.368±.011	0 (60 frozen)
untrained Quadratic control	0.352±.010	0.349±.010	0 (164 frozen)
rotation-plane head (isometric control)	0.405±.008	0.402±.008	60

The readout-only variant remains trainable and reaches 0.412 P@5, while re-uploading supplies the remaining 0.016. The rotation-plane result is a closer structural comparison than the inert fidelity control, and its 0.023 gap motivates the equal-budget analysis in Appendix J.3.

Table 7: Report generation with frozen medical LVLMs given top-5 retrieved reports as context (5 seeds; std $\leq$ 0.004 for adapted retrievers, ViT Only deterministic under greedy decoding; LoRA r=16 tracks Adapter within 0.003 and is omitted). B-4, R-L, MTR denote BLEU-4, ROUGE-L, METEOR. Marks and tints as in Table 1.

Dataset	Retriever	LLaVA-Med (7B)				CheXagent (8B)			
		B-4	R-L	MTR	CIDEr	B-4	R-L	MTR	CIDEr
IU X-Ray	ViT Only	0.134	0.298	0.167	0.241	0.147	0.312	0.178	0.263
	Adapter	0.156	0.332	0.189	**0.298**	0.171	0.348	0.203	**0.321**
	QuPID	**0.163**	**0.341**	**0.194**	0.294	**0.178**	**0.357**	**0.208**	0.318
MIMIC-CXR	ViT Only	0.098	0.247	0.138	0.168	0.108	0.261	0.149	0.184
	Adapter	0.121	0.281	0.163	**0.221**	0.134	0.298	0.176	**0.242**
	QuPID	**0.127**	**0.289**	**0.168**	0.218	**0.141**	**0.307**	**0.182**	0.239

The ordering depends on the metric: QuPID improves BLEU-4, ROUGE-L and METEOR over Adapter, whereas Adapter retains a small CIDEr lead. Retrieval therefore changes report quality without producing a uniform gain across lexical measures.

Table 8: Clinical efficacy proxies on MIMIC-CXR (5 seeds; std ≤ 0.005): CheXbert micro-F1 and RadGraph-F1 of generated reports. †, marks and tints as in Table 1.

Retriever	LLaVA-Med (7B)		CheXagent (8B)	
	CheXbert-F1	RadGraph-F1	CheXbert-F1	RadGraph-F1
ViT Only	0.298	0.181	0.321	0.194
Adapter	0.334	0.203	0.356	0.217
QuPID	**0.352**†	**0.214**†	**0.373**†	**0.229**†

The finding-based proxies favor QuPID for both generators, extending the comparison beyond surface wording. These scores measure agreement between extracted findings and references, without a clinical-outcome endpoint.

Table 9: Computational cost with the shared frozen ViT-L/16 (ChestX-ray14, batch 32, single A100, full budget). Latency: readout and top-k search over cached archive readouts at batch size 1, excluding the shared ViT pass. Memory: peak allocation during adaptation.

Method	#Params	s/epoch	Lat. (ms)	Mem. (MB)
ViT Only	0	n/a	6.2	3,024
Adapter	525K	44.6	7.1	3,728
LoRA (r=16)	1.57M	52.3	7.0	3,810
QuPID	60	71.3	7.6	3,582

Table 10: Embedding diagnostics (5 seeds, adapted-entry std ≤ 0.009; frozen ViT Only is deterministic). Jac@10: mean query–neighbor label-set Jaccard. MURA Jac@10 is derived from P@10 in Table 11 by the singleton-label identity; n/a: no corresponding P@10 is reported. $\mathrm{Sil}^{\mathrm{SL}}$: silhouette on the single-label subset.

Method	ChestX-ray14		MURA	
	Jac@10	$\mathrm{Sil}^{\mathrm{SL}}$	Jac@10	$\mathrm{Sil}^{\mathrm{SL}}$
ViT Only	0.218	0.182	0.456	0.157
Adapter	0.241	0.228	0.548	0.213
Adapter-XL	0.249	0.239	0.558	0.221
Quadratic	0.238	0.224	n/a	0.208
QuPID	**0.264**	**0.273**	**0.581**	**0.249**

The cost table separates training from cached retrieval: fewer trainable parameters do not imply a faster epoch, and the reported latency excludes the shared encoder. The neighboring geometry table describes representation structure rather than deployment cost.

Table 11: Retrieval metrics (5-seed means; adapted-entry stds within ± 0.009, omitted for space). Marks and tints as in Table 1.

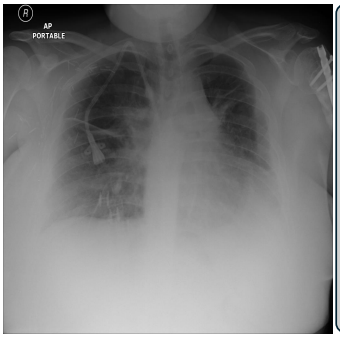
Method	P@5	P@10	MAP@10	MRR	NDCG@5	NDCG@10
ChestX-ray14						
ViT Only	0.312	0.287	0.278	0.363	0.325	0.307
Linear Head	0.378	0.351	0.341	0.434	0.389	0.368
Adapter	0.401	0.375	0.368	0.462	0.415	0.396
Adapter-L	0.412	0.386	0.372	0.465	0.421	0.401
Adapter-XL	0.414	0.388	0.387	**0.468**	0.433	0.414
QuPID	**0.428**	**0.404**	**0.399**	0.458	**0.443**	**0.425**
MURA						
ViT Only	0.481	0.456	0.451	0.529	0.494	0.475
Linear Head	0.549	0.523	0.513	0.594	0.562	0.541
Adapter	0.573	0.548	0.546	0.618	0.587	0.567
Adapter-L	0.581	0.555	0.556	0.625	0.598	0.580
Adapter-XL	0.584	0.558	0.562	0.632	0.604	0.584
QuPID	**0.601**	**0.581**	**0.579**	**0.649**	**0.618**	**0.599**

The rows distinguish the gains: QuPID leads precision and NDCG, while ChestX-ray14 MRR favors Adapter-XL (0.468 against 0.458).

F QUALITATIVE EXAMPLES

Figures 10a and 10b show representative retrieval-grounded reports, each displaying a query image and the report generated from QuPID-retrieved context.

The MURA example illustrates retrieval-grounded reporting beyond chest imaging; the quantitative generation evaluation of Section 4.1 covers the chest datasets, where paired report corpora and specialized generators exist.



QuPID Reports:

The **trachea** is midline. Cardiac silhouette is normal in size. **Bilateral lower lobe infiltrates** with reticular and **airspace opacities** are identified, more prominent in the **lower zones**. The pattern is consistent with **pneumonia** versus **pulmonary edema**. No **pleural effusion** or **pneumothorax**. The **costophrenic angles** are sharp. Visualized osseous structures are unremarkable. Lucent pulmonary parenchyma consistent with **emphysema** appears unchanged from prior. Clinical correlation is recommended to differentiate infectious versus cardiogenic etiology. Follow-up imaging after treatment is advised to ensure resolution.

(a) Chest X-ray query.



QuPID Reports:

Lateral view of the **forearm** demonstrates a **transverse fracture** through the mid-shaft of the **radius** with approximately 2-3mm displacement. No **comminution**. The **ulna** is intact with preserved cortical margins. The **radioulnar joint** alignment is maintained without **dislocation**. Mild **soft tissue swelling** is present overlying the fracture site. The visualized **elbow** and **wrist** articulations appear unremarkable. Recommend **orthopedic consultation** for closed reduction with long-arm casting versus surgical fixation. Follow-up radiographs in 1-2 weeks to assess alignment and at 4-6 weeks to evaluate for **callus formation** and healing.

(b) Forearm radiograph query.

Figure 10: Retrieval-grounded reports: each panel shows the query image and the report a frozen LVLM generates from QuPID's top-5 retrieved cases. The retrieved cases in (a) are pathologically consistent and the generator narrows the differential to pneumonia versus pulmonary edema; in (b) they support a displaced transverse fracture of the radius.

These cases illustrate how retrieved evidence enters a generated report. They complement the aggregate evaluations above, but two selected examples cannot estimate how often findings are supported.

G ROBUSTNESS AND TRAINABILITY DIAGNOSTICS

Sensitivity of amplitude encoding to feature perturbations. Amplitude encoding renormalizes the feature vector, raising the concern that small perturbations of the classical features could alter all amplitudes and destabilize retrieval. Table 12 adds zero-mean Gaussian noise scaled to the root-mean-square feature magnitude (perturbation $\sigma \cdot \mathrm{RMS}(\mathbf{h}) \cdot \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I)$) to the frozen features before L2 normalization and measures retrieval quality. QuPID degrades at a rate comparable to the classical adapter across the tested range, indicating that the encoding does not disproportionately amplify feature noise at the operating dimensionality; the potential concern is not supported by the measurements at this scale.

Gradient variance versus depth. Table 13 reports the empirical variance of parameter-shift gradients of the contrastive loss at initialization ($\boldsymbol{\theta}, \boldsymbol{\alpha} \sim \mathcal{U}(-\pi, \pi)$; variance taken over parameters, 5 seeds, and 20 mini-batches) as circuit depth grows; $L = 12$ is included purely as a stress test beyond the $L \leq 6$ design range. Because all asymptotic barren plateau statements concern scaling in n_q, which is fixed at 10 here, the informative comparison is the reference decay rate: the observed factor

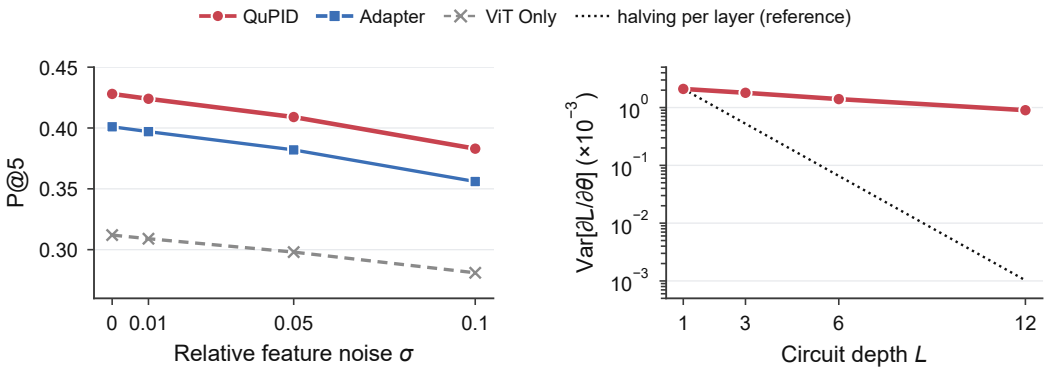
Table 12: Retrieval under feature perturbation on ChestX-ray14 (P@5, 5 seeds; std ≤ 0.009). Relative noise σ is applied before L2 normalization. Marks and tints as in Table 1.

Method	σ=0	σ=0.01	σ=0.05	σ=0.1
ViT Only	0.312	0.309	0.298	0.281
Adapter	0.401	0.397	0.382	0.356
QuPID	**0.428**	**0.424**	**0.409**	**0.383**

Table 13: Variance of parameter-shift gradients at initialization ($\times 10^{-3}$, ChestX-ray14, 5 seeds).

	L=1	L=3	L=6	L=12
$\mathrm{Var}[\partial\mathcal{L}/\partial\theta]$	2.1	1.8	1.4	0.9

The two diagnostics separate input sensitivity from optimization signal. A drop under feature noise concerns retrieval robustness; nonzero gradient variance at initialization addresses a different question and does not certify optimization success.



(a) Feature noise.

(b) Gradient variance against depth.

Figure 11: Diagnostics for feature sensitivity and gradient decay in this architecture (ChestX-ray14, 5 seeds); the design range in (b) is $L \leq 6$.

of 2.3 from $L = 1$ to $L = 12$ is far from the multiplicative collapse per layer that global observables exhibit, consistent with the local one- and two-body readout; at the default $L = 3$ the gradient signal sits well within the trainable range.

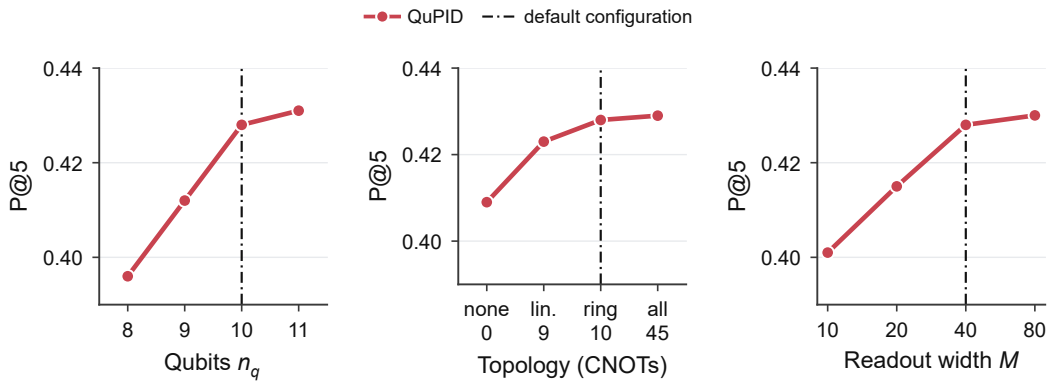
Two further checks examine qubit-count dependence and parameter updates. Varying n_q from 6 to 12 with the input construction of Table 14, the gradient variance at initialization moves from 3.9 to 2.2 ($\times 10^{-3}$), a factor of 1.8, against 64 for a 2^{-n_q} collapse. After adaptation, the mean absolute parameter change is 0.68 rad for the rotation angles and 0.41 for the re-uploading scales, 98% of the 60 parameters move by more than 0.05 rad, and none stays within 0.01 of its initialization, so the readout is not an untrained random feature map.

H Circuit Geometry, Sensitivity and Transfer

Section 4.2 removes one design choice at a time inside a fixed circuit. This appendix instead varies the circuit itself, then asks whether the reported gain depends on hyperparameters or on evaluating at the site the model was adapted on. These studies evaluate potential sensitivities of the method and report both saturation and performance gains.

H.1 Circuit design: qubits, entangling gates, and observables

Figure 12 plots all three sweeps and Tables 14 to 16 give the underlying numbers. Three readings follow. Qubit count is not a free parameter in our design, since amplitude encoding ties it to the feature dimension; Table 14 varies it by reducing the backbone feature dimension by projection or increasing it by concatenating features from two backbones, and the curve saturates once the native 1024 dimensions are used, so the $+0.003$ available at $n_q = 11$ costs a doubling of state-vector width for a gain within the observed variation across seeds. Depth is the one axis on which we deliberately stay small: at fixed qubit count, deeper re-uploading of high-dimensional inputs is known to degrade predictive performance toward chance (Wang et al., 2025e), and our depth ablation saturates by $L = 3$, so the design is wide in its readout ($M = 40$ observables) rather



(a) Qubit count. (b) Entangling topology. (c) Readout width M.

Figure 12: Three circuit-design axes on ChestX-ray14 (P@5, 5 seeds); dash-dotted lines mark the default configuration, and (b) lists CNOTs per layer under each topology.

Table 14: Qubit count. n_q is fixed by the feature dimension, varied here by projection or concatenation. Marks and tints as in Table 1.

Input	n_q	P@5
PCA-256	8	0.396
PCA-512	9	0.412
native 1024	10	0.428
concat 2048	11	**0.431**

Table 15: Entangling topology at n_q=10 and the CNOT count each one needs per layer. Marks and tints as in Table 1.

Topology	CNOT	P@5
none	0	0.409
linear	9	0.423
ring	10	0.428
all-to-all	45	**0.429**

Table 16: Readout width; observable families are added in the listed order. Marks and tints follow Table 1.

Observables	M	P@5
Z	10	0.401
Z, X	20	0.415
$+ZZ, XX$	40	0.428
$+Y, YY$	80	**0.430**

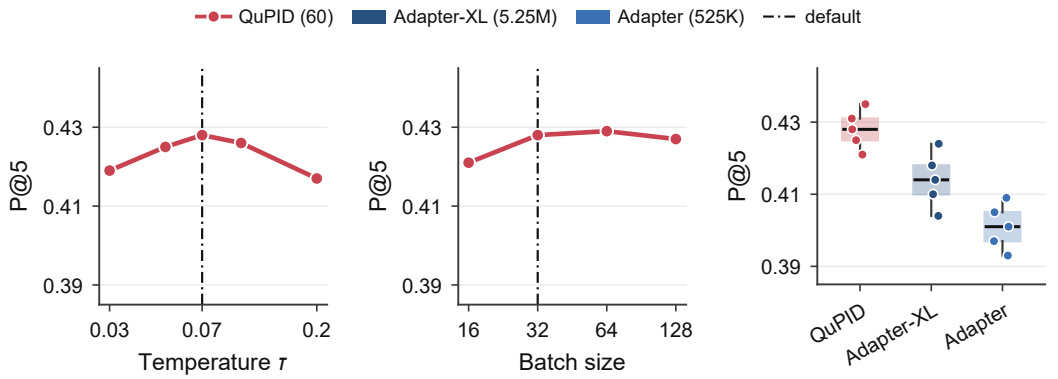
The design sweep changes feature access, connectivity and measurement width separately. Additional qubits also change the input representation, whereas additional observables change the readout; their gains should not be attributed to one common source of capacity.

than deep in its encoding. Entanglement contributes at low gate cost (Table 15): removing CNOTs entirely costs 0.019 P@5, while ring connectivity recovers all but 0.001 of what all-to-all achieves using 10 gates per layer instead of 45. This is the concrete answer to whether the correlations introduced by the circuit contribute to retrieval performance, and it is also why we did not pursue richer topologies. We also evaluate the choice of rotation axis: replacing R_Y by R_X gives 0.421 P@5 and replacing it by R_Z gives 0.404, while adding an R_Z layer after R_Y (90 parameters) gives 0.429; the default achieves a nearby mean with fewer parameters. Simulation cost follows the qubit sweep as the exponent predicts: relative to $n_q = 10$, per-epoch time is 0.41, 0.58, 1.00, 2.9 and 10.4 at $n_q = 6, 8, 10, 12, 14$ and peak memory 0.62, 0.71, 1.00, 1.9 and 5.6, which is why n_q is set by the feature dimension rather than tuned upward. Readout expansion beyond two-body terms gains only 0.002 P@5 (Table 16), consistent with the factorization view of Section 3.3. For real-valued states, odd-Y Pauli strings have zero expectation, whereas even-Y terms such as Y_jY_k can be nonzero.

H.2 Hyperparameter sensitivity

Because all hyperparameters were fixed a priori on a disjoint pilot subset (Section 4), we examine whether the chosen settings lie near a narrow performance peak. Figure 13 examines this sensitivity together with variability across seeds. Panels (a) and (b) sweep the two settings the objective is most sensitive to and find plateaus in both: P@5 varies by at most 0.011 over a sevenfold range of temperature and by 0.008 over an eightfold range of batch size, with the defaults inside the flat region rather than at its maximum. We take this as evidence that the reported results do not rely on additional tuning, not as evidence that the method is insensitive to every hyperparameter.

Panel (c) makes the seed-level spread explicit for the comparison underlying the main empirical claim. The QuPID and Adapter-XL distributions are close enough that individual seeds interleave, which is exactly why the significance protocol of Section 4 operates on per-query paired resamples rather than on the seed means; the mean gap alone would overstate their separation.



(a) Temperature τ. (b) Batch size. (c) Per-seed P@5 at the full budget.

Figure 13: Sensitivity and spread (ChestX-ray14, 5 seeds); dash-dotted lines mark default settings.

H.3 Cross-site transfer

Table 17: Cross-site zero-shot transfer (P@5, 5 seeds; std $\leq$ 0.009). Modules are adapted on ChestX-ray14 and evaluated without further training. Marks and tints as in Table 1, by column.

Evaluation	frozen	cls.	QuPID
ChestX-ray14 (in-domain)	0.312	0.401	**0.428**
CheXpert (zero-shot)	0.289	0.301	**0.318**

Table 17 examines the open question of whether a module adapted at one site helps at another. Adapting on ChestX-ray14 and evaluating on CheXpert without further training, both modules stay above the frozen baseline: QuPID gains 0.029 and the reported adapter 0.012, leaving a 0.017 margin between them. We read this as encouraging rather than settled. The margin between adapted modules falls from 0.027 to 0.017, retaining about 63% of its in-domain value; the gains over frozen features fall more sharply. A single site pair cannot establish that this ordering holds in general. Cross-site adaptation therefore remains future work, informed by this measurement.

I Noise-Aware Simulation

Every result in the main text comes from a noiseless statevector simulator, the prevailing convention in the QML literature (Appendix A.6). The separate noise study evaluates an optional quantum-hardware realization of the learned readout. It measures sensitivity to the specified channels and sampling budgets; the GPU deployment uses exact statevector expectations. This appendix therefore reports two analyses, a post-training sensitivity study under standard noise channels and a shot-budget calculation for finite-sample readout estimation. Both are simulator-level diagnostics, not hardware benchmarks, and support no robustness claim about physical devices.

I.1 Noise channels

On hardware the state is described by a density matrix ρ rather than a state vector, and imperfections act as completely positive trace-preserving maps $\mathcal{E}(\rho) = \sum_k K_k \rho K_k^\dagger$ with $\sum_k K_k^\dagger K_k = I$. We evaluate the four channels standard in this literature, applied independently to each qubit after every circuit layer with strength p:

$$\text{depolarizing:} \quad \mathcal{E}(\rho) = (1-p)\rho + \tfrac{p}{3}\big(X\rho X + Y\rho Y + Z\rho Z\big), \tag{9}$$

$$\text{bit flip:} \quad \mathcal{E}(\rho) = (1-p)\rho + p\, X\rho X, \tag{10}$$

$$\text{phase flip:} \quad \mathcal{E}(\rho) = (1-p)\rho + p\, Z\rho Z, \tag{11}$$

$$\text{amplitude damping:} \quad \mathcal{E}(\rho) = K_0\rho K_0^\dagger + K_1\rho K_1^\dagger, \tag{12}$$

with $K_0 = \mathrm{diag}(1, \sqrt{1-p})$ and $K_1 = \sqrt{p}\,|0\rangle\langle 1|$. Depolarizing noise models unstructured gate error, the two flip channels model Pauli errors in the computational and phase bases, and amplitude

damping models energy relaxation toward $|0\rangle$; unlike the first three it is non-unital, so it does not preserve the maximally mixed state and can bias expectations rather than merely shrinking them.

Two structural remarks explain what to expect. The evaluated Pauli channels act multiplicatively on Pauli components at the channel location, $\langle P\rangle \mapsto (1 - c_P p)\langle P\rangle$, where c_P depends on the observable as well as the channel; the readout direction can therefore change at first order. Because the cosine of Section 3.3 is scale-invariant, a common multiplicative factor cancels. Observable-dependent factors and noise interleaved with noncommuting gates need not cancel, so the ranking effect must be measured. Amplitude damping has no such protection, since it moves the readout rather than scaling it. The measured pattern below is consistent with this analysis, and the norm floor c of Proposition 3 is the quantity that controls how much contraction cosine similarity can tolerate before the similarity becomes ill-conditioned.

I.2 Post-training noise sensitivity

We isolate inference-time sensitivity: the model is trained once on the noiseless simulator, its 60 parameters are frozen, and only evaluation is perturbed. Archive and query readouts pass through the same channel, so the study measures degradation of the retrieval geometry rather than a train-test mismatch. Table 18 reports P@5 on ChestX-ray14 across channel strengths, using the density-matrix simulator (`default.mixed`) with the noise hook of Listing 1.

Table 18: Post-training noise sensitivity on ChestX-ray14 (P@5, 5 seeds, std ≤ 0.007). The trained model is frozen and noise is injected only at evaluation, per qubit after every layer. The frozen-baseline P@5 of 0.312 marks zero accuracy gain over frozen retrieval.

Channel	p=0	p=0.001	p=0.005	p=0.01	p=0.05
Depolarizing	0.428	0.428	0.426	0.423	0.404
Bit flip	0.428	0.427	0.424	0.419	0.392
Phase flip	0.428	0.428	0.427	0.426	0.418
Amplitude damping	0.428	0.427	0.423	0.417	0.386

Degradation is gradual across all four channels: at $p = 0.01$, a per-qubit per-layer error rate comparable to current two-qubit gate fidelities, the largest drop is 0.011 P@5, and even at the deliberately pessimistic $p = 0.05$ the adapted readout stays well above the frozen baseline of 0.312. Phase flip causes the smallest performance degradation in this experiment, while amplitude damping and bit flip cause the largest. Non-unital damping can bias expectations, and bit flips alter computational-basis populations; commutation with final observables alone does not determine the effect of noise inserted between circuit layers. We read these numbers as a sensitivity diagnostic showing that the learned geometry is not excessively sensitive to the modeled perturbations, not as evidence of device-level robustness: the study omits crosstalk, coherent and correlated errors, readout error, and hardware-specific gate decomposition, all of which a real device would add.

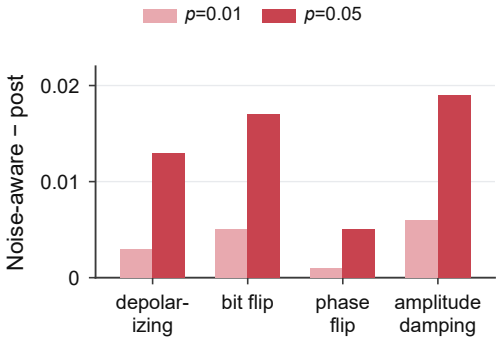


Figure 14: Gain of noise-aware training over post-training injection at two channel strengths (P@5, ChestX-ray14, 5 seeds; absolute values in Table 19).

I.3 NOISE-AWARE TRAINING

Figure 14 plots the performance recovery that Table 19 tabulates. The study above freezes a model trained without noise, which answers whether the learned geometry *tolerates* noise but not whether the 60 parameters can be *learned* inside a noisy channel. The latter is the question that matters for eventual on-device adaptation, and it is not implied by the former: noise could in principle flatten the loss surface enough to stall optimization even where it barely perturbs a fixed readout. Table 19 repeats end-to-end adaptation with the same channel strength during training and evaluation.

Table 19: Noise-aware training versus post-training injection (ChestX-ray14, P@5, 5 seeds, std ≤ 0.008). Each entry trains and evaluates at the same channel strength; *post* repeats the frozen-model numbers of Table 18 for comparison. Δ is the recovery from training through the channel.

Channel	$p = 0.01$			$p = 0.05$		
	post	noise-aware	Δ	post	noise-aware	Δ
Depolarizing	0.423	0.426	$+0.003$	0.404	0.417	$+0.013$
Bit flip	0.419	0.424	$+0.005$	0.392	0.409	$+0.017$
Phase flip	0.426	0.427	$+0.001$	0.418	0.423	$+0.005$
Amplitude damping	0.417	0.423	$+0.006$	0.386	0.405	$+0.019$

Two patterns are worth naming. Training through the channel recovers part of the loss in every case, and the recovery is largest exactly where the post-training performance degradation was largest, which is what one expects if the circuit is re-allocating its rotations toward observables the channel leaves better conditioned rather than merely averaging noise away. The recovery is nonetheless partial: no configuration returns to the noiseless 0.428, so noise-aware training mitigates rather than removes the cost, and we do not claim that a noisy channel is harmless or that it acts as a useful regularizer. Gradients remain exact here, since the parameter-shift rule applies to noisy-channel expectations as it does to the unitary case; what changes is the differentiated value, not the estimator.

I.4 SHOT BUDGET FOR HARDWARE READOUT

Condition on the learned circuit and a fixed collection of K inputs. For each input i and observable m, estimate $z_{i,m}$ by the mean $\hat{z}_{i,m}$ of S independent shots $Y_s \in \{-1, +1\}$ with expectation $z_{i,m}$. Hoeffding's inequality for this range of length 2 gives

$$\Pr\big(|\hat{z}_{i,m} - z_{i,m}| > \varepsilon\big) \leq 2\exp(-S\varepsilon^2/2).$$

A union bound over the KM estimates yields $\max_{i,m} |\hat{z}_{i,m} - z_{i,m}| \leq \varepsilon$ with probability at least $1 - \delta$ whenever $S \geq 2\log(2KM/\delta)/\varepsilon^2$. Independence between different observables or inputs is not needed for this union bound. Assume the exact readouts in this collection satisfy $\|\mathbf{z}_i\| \geq c > 0$. On the event above, let $r = \sqrt{M}\varepsilon < c$, so $\|\hat{\mathbf{z}}_i - \mathbf{z}_i\| \leq r$ and $\hat{\mathbf{z}}_i \neq 0$. The exact normalization inequality

$$\left\| \frac{\hat{\mathbf{z}}_i}{\|\hat{\mathbf{z}}_i\|} - \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|} \right\| \leq \frac{2\|\hat{\mathbf{z}}_i - \mathbf{z}_i\|}{\|\mathbf{z}_i\|} \leq \frac{2r}{c}$$

implies that each query–candidate cosine changes by at most $4r/c$. The difference between two such scores changes by at most $8r/c$. For $0 < \gamma < 2$, choose $\varepsilon = c\gamma/(8\sqrt{M})$, which also ensures $r < c$. Every comparison whose exact score gap is greater than γ then retains its ordering with probability at least $1 - \delta$ if

$$S \;\geq\; \frac{128\, M \log(2KM/\delta)}{(c\gamma)^2}. \tag{13}$$

For one fixed query and two candidates, take $K = 3$. For one query and an archive of N items, take $K = N + 1$ to cover all their readouts simultaneously; the same event protects every candidate comparison with a gap greater than γ. The bound does not protect smaller gaps. Under the separate-observable sampling protocol, caching the archive costs NMS shots once and each new query costs MS shots at fixed S. For a simultaneous archive-wide guarantee, however, the sufficient S in equation 13 depends logarithmically on N. This is a sufficient worst-case readout guarantee, conditional on the norm floor and score gaps, not a numerical prediction of average P@5. Without specifying c, γ, δ and K, it does not imply $S = 10^4$. It also assumes noiseless circuit execution; gate noise and hardware latency require separate analysis.

I.5 FINITE-SHOT READOUT

Separately from channel noise, a device returns each expectation only up to sampling error. Table 20 replaces exact expectations with S-sample estimates for both query and archive readouts, leaving the circuit noiseless so that the two error sources are evaluated separately.

Table 20: Finite-shot readout on ChestX-ray14 (5 seeds, std ≤ 0.009). Each of the $M = 40$ observables is estimated from S shots for query and archive alike; $S = \infty$ denotes exact expectations.

Shots per observable	S=10^2	S=10^3	S=10^4	S=∞
P@5	0.389	0.418	0.426	0.428
NDCG@10	0.386	0.415	0.423	0.425

Empirically, retrieval is within 0.002 P@5 of the exact-expectation reference at $S = 10^4$, while $S = 10^2$ gives a larger loss. This average-metric result does not establish the simultaneous ranking guarantee of Appendix I.4, which depends on the norm floor, score gaps, collection size and failure probability. Classical adapters incur no measurement-shot cost. Table 9 reports simulator timings and cannot establish device latency at 10^4 shots per observable; the experiment here isolates finite-sampling sensitivity rather than hardware feasibility.

```python
import pennylane as qml

def apply_noise_channel(noise, wires):
    """Per-qubit noise applied after each QuPID layer (default.mixed)."""
    channel, p = noise
    for w in wires:
        if channel == "depolarizing":
            qml.DepolarizingChannel(p, wires=w)
        elif channel == "bit_flip":
            qml.BitFlip(p, wires=w)
        elif channel == "phase_flip":
            qml.PhaseFlip(p, wires=w)
        elif channel == "amplitude_damping":
            qml.AmplitudeDamping(p, wires=w)
        else:
            raise ValueError(f"unknown channel: {channel}")
```

Listing 2: Noise hook used for the density-matrix study; `noise` is a (channel, strength) pair and the function is the `apply_noise_channel` referenced in Listing 1.

J WHERE THE ADVANTAGE COMES FROM

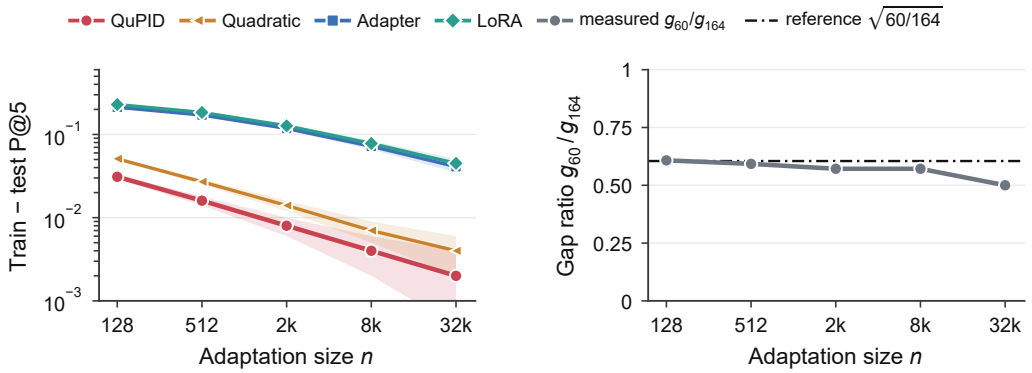
Section 4.1 reports how retrieval performance varies with adaptation size. A performance curve alone does not identify the mechanism: the classical baselines may be undertuned at small n, or another classical module at the same budget may close the gap. This appendix separates them. It examines train–test retrieval gaps, retunes the reported baseline configurations, compares further classical modules at our budget, and scales a spectrum-sampled surrogate far past that budget.

J.1 TRAIN–TEST RETRIEVAL GAPS

Table 21: Training P@5 minus held-out P@5 on ChestX-ray14 after label-free adaptation (5 seeds; std ≤ 0.006 for the two large-budget families and ≤ 0.002 for the two small ones). This retrieval diagnostic is distinct from the contrastive-risk gap bounded by Proposition 3. Lower is better; marks and tints as in Table 1.

Method	p	n=128	512	2,048	8,192	32,768
QuPID	60	**0.031**	**0.016**	**0.008**	**0.004**	**0.002**
Quadratic	164	0.051	0.027	0.014	0.007	0.004
Adapter	525K	0.212	0.171	0.118	0.072	0.041
LoRA (r=16)	1.57M	0.229	0.184	0.127	0.078	0.045

Table 21 and Figures 15(a) and 3(a) report the gap itself. The gaps range from 0.031 to 0.229 at $n = 128$ and from 0.002 to 0.045 at the full budget, and increase with parameter budget among these four families at every adaptation size. Plotted against $\sqrt{p/n}$, the two small-budget families follow



(a) Gap against n.

(b) Gap ratio and budget reference.

Figure 15: Train–test P@5 gaps (ChestX-ray14, 5 seeds; bands one standard deviation). The families are ordered by parameter budget at every n. The line at $\sqrt{60/164}$ in (b) is a parameter-count reference, not a prediction for the ratio of the QuPID gap g_{60} to the Quadratic gap g_{164}.

a similar trend, while the two large-budget families fall below the slope-one guide in Figure 3(a). This is a descriptive comparison of retrieval gaps, not a test of whether the contrastive-risk bound is tight or vacuous. Two observations summarize the pattern. First, the smaller models have smaller observed P@5 gaps throughout this sweep. A parameter-count upper bound does not establish this ordering of realized gaps. Second, the two small-budget families have similar empirical scaling with adaptation size. The measured gap ratio between the 60-parameter and 164-parameter readouts is 0.61, 0.59, 0.57, 0.57, and 0.50 across the sweep, alongside the reference $\sqrt{60/164} = 0.60$ in Figure 15(b). We state the scope plainly. The reference omits class-dependent constants and concerns a different loss from P@5. Its numerical proximity to the measured ratios therefore supplies no theorem-based prediction for retrieval. These observations do not isolate capacity as the cause: the four-projection Quadratic head has an information bottleneck. The rotation-plane train–test sweep in Appendix J.3 provides the separate equal-parameter comparison; it is not part of this figure.

J.2 Capacity- and regularization-matched baselines

Table 22: Reported baseline retuning sweep on ChestX-ray14 (P@5, 5 seeds). *fixed* uses the a priori protocol; *tuned* selects width, weight decay, dropout, and patience at each n by held-out contrastive loss. QuPID keeps its configuration fixed. Margins apply to this sweep and exclude the additional controls in Table 23. Marks and tints as in Table 1, by column.

n	**sweep baseline**		**QuPID**	**margin**	
	fixed	tuned		fixed	tuned
128	0.316	0.331	0.352	+0.036	+0.021
512	0.347	0.361	0.401	+0.054	+0.040
2,048	0.378	0.384	0.416	+0.038	+0.032
8,192	0.401	0.405	0.424	+0.023	+0.019
32,768	0.414	0.416	0.428	+0.014	+0.012

Adaptation data size changes the useful model capacity and regularization, motivating a separate validation-selected adapter sweep. Table 22 tests how retuning the reported configurations changes the comparison. At each adaptation size we search these configurations over bottleneck width, weight decay, dropout, and early-stopping patience, and select on a held-out contrastive validation split; QuPID keeps its fixed configuration. Tuning improves baseline performance most at small adaptation sizes: it recovers 0.015 P@5 at $n = 128$ and 0.014 at $n = 512$, against 0.002 at the full budget. The margin at 512 falls by roughly a quarter, from $+0.054$ to $+0.040$, and Figure 2(c) shows the two margin curves keeping the same shape, peaked at 512 and positive at every n. This sweep does not include every classical control: at $n = 512$, the rotation-plane head in Table 23 scores 0.386, above either baseline column here, leaving a smaller QuPID margin of $+0.015$. The selected configurations are themselves informative: at $n = 128$ the search selects the narrowest bottleneck and largest weight decay, then selects a wider, less regularized configuration at 512. Concretely, the grid covers bottleneck widths $\{32, 64, 128, 256, 512\}$, weight decay $\{0, 0.01, 0.05, 0.2\}$, dropout

$\{0, 0.1, 0.3\}$ and early-stopping patience $\{5, 10, 20\}$, and the selection moves monotonically with the adaptation size: width 32 with decay 0.2 and dropout 0.3 at $n = 128$, width 64 with decay 0.05 at 512, width 128 at 2,048, and the widest setting with decay 0.01 and no dropout at both remaining sizes. The selected patience likewise increases from 5 to 20 as data grows. The selected configurations respond to scarcity by becoming smaller and more regularized; their retrieval scores remain below QuPID within this sweep. Proposition 3 does not guarantee this ordering. The comparison fixes both the validation criterion and the search space. Only the tuned adapter uses held-out contrastive loss to select among the listed settings; QuPID retains its a priori configuration. The finite search space defines the scope of this comparison.

Selected adapter configurations. The tuned entries in Table 22 use a two-layer residual bottleneck adapter. The selected width, weight decay, dropout and early-stopping patience at each adaptation size are:

n	Width	Decay	Dropout	Patience
128	32	.20	.30	5
512	64	.05	.10	10
2,048	128	.05	.10	10
8,192	512	.01	.00	20
32,768	512	.01	.00	20

Training follows Table 4; held-out contrastive loss selects the settings. Appendix M.1 specifies the separate rotation-plane tuning budget.

J.3 Other classical modules at the same budget

Table 23: A range of classical adaptation modules at or near the 60-parameter budget (P@5, ChestX-ray14, 5 seeds; std ≤ 0.009). Each is trained with the label-free objective and schedule of Section 4; the frozen encoder scores 0.312. Marks and tints as in Table 1.

Module	p	**full budget**	n=512
Diagonal gain on 60 leading PCA directions	60	0.331	0.322
Random Fourier features	64	0.319	0.311
Tensor-train head over the ten blocks, bond dimension 2	64	0.361	0.348
Rotation-plane isometry	60	0.405	0.386
Quadratic head	164	0.392	0.372
Bias-only shift	1,024	0.352	0.337
LoRA (r=1), one projection	2,048	0.386	0.349
QuPID	60	**0.428**	**0.401**

Table 23 answers a question the parameter-matched controls of Table 1 leave open: whether some other classical module at the same budget, rather than the two we chose, would close the gap. The comparison covers the natural candidates: a diagonal rescaling, a tensor-train contraction over the same ten two-dimensional blocks the circuit acts on, bias-only and rank-one updates, and the isometric rotation-plane head. Under these default configurations, the readout has the highest mean P@5 at both adaptation sizes. The default isometric control comes closest. At full budget and at 512 examples, QuPID reaches P@5 of 0.428 and 0.401, versus 0.405 and 0.386, giving margins $+0.023$ and $+0.015$. Its full-to-512 decrease is 0.027, compared with 0.019 for the control. The full sweep below reports accuracy ordering; Appendix M.1 reports tuned-control scores, and Appendix M.2 uncertainty at the default endpoints.

Rotation-plane adaptation-size sweep. The following measurements use the fixed rotation-plane configuration of Appendix D at all five adaptation sizes. The last column is training P@5 minus held-out P@5 for that control.

n	QuPID P@5	Rotation P@5	Rotation train–test gap
128	0.352	0.337	0.039
512	0.401	0.386	0.023
2,048	0.416	0.397	0.012
8,192	0.424	0.403	0.008
32,768	0.428	0.405	0.004

QuPID leads the P@5 means by 0.015, 0.015, 0.019, 0.021 and 0.023 as n grows. The rotation-plane train–test gap falls from 0.039 to 0.004. Higher held-out accuracy and a smaller train–test gap are separate observations; the parameter-count bound does not determine either ordering.

J.4 Stronger classical surrogates

Appendix J.3 holds the classical budget at 60 parameters. This subsection removes that restriction and asks how large and how well tuned a classical surrogate must be before it reproduces the readout. The surrogates are ordered by how much of the circuit they are given.

Spectrum-sampled random features. Proposition 2 makes each readout component a trigonometric polynomial in the re-uploaded projections, with frequencies in signed sums of the rates $\alpha_j^{(l)} 2^{j-1}$. The surrogate draws D frequency vectors from that support, forms features $\cos(\boldsymbol{\omega}_d^\top \mathbf{s}(\mathbf{x}) + b_d)$ and learns a $40 \times D$ output map, so $p = 40D$. This is the construction Landman et al. (2023) propose for variational models, instantiated on our encoding rather than a generic kernel.

Structure-matched heads. Two further surrogates receive the algebraic structure of the readout without the circuit. The first applies a learned orthogonal map built from Givens rotations and measures traceless observables with ± 1 spectrum, which is the readout of Section 3.3 with re-uploading removed and the circuit constraint lifted. The second learns 40 unconstrained symmetric forms of rank r. The rotation-plane head of Appendix J.3 is its rank-one special case.

Budget. Every surrogate receives more than QuPID: 100 validation trials over learning rate, batch size, width and regularization, against the 30 trials of Appendix M.1, and $1.5\times$ its adaptation wall-clock. QuPID keeps the a priori configuration of Table 4.

Table 24: Classical surrogates given more parameters and more tuning than QuPID (ChestX-ray14, P@5, five adaptation subsets crossed with five initializations). Δ is the paired mean difference in favour of QuPID, with a Holm-corrected 95% cluster-bootstrap interval over subsets and patients. Marks and tints as in Table 1.

Surrogate	p	**P@5**		**Δ in favour of QuPID**	
		n=512	full	at 512	at full
Spectrum features, D=16	640	0.318	0.341	+0.083	+0.087
Spectrum features, D=128	5,120	0.339	0.372	+0.062	+0.056
Spectrum features, D=1,024	41,000	0.357	0.396	+0.044	+0.032
Spectrum features, D=8,192	328,000	0.362	0.407	+0.039 [.028,.050]	+0.021 [.013,.029]
Givens rotations, Pauli-type observables	60	0.389	0.409	+0.012	+0.019
Givens rotations, Pauli-type observables	600	0.392	0.417	+0.009 [−.002,.020]	+0.011 [.004,.018]
Unconstrained quadratic forms, r=4	6,600	0.377	0.412	+0.024	+0.016
Rotation-plane head, tuned	60	0.394	0.414	+0.007 [−.004,.018]	+0.014 [.006,.022]
QuPID	**60**	**0.401**	**0.428**	n/a	n/a

Paired analysis. Each surrogate and QuPID are fitted on the same five adaptation subsets crossed with the same five initializations, giving 25 paired runs per budget and no unpaired comparison. The estimand is the mean per-query P@5 difference over subsets, initializations and held-out patients. Intervals resample subset draws with nested initializations and patient clusters, and win rates count queries on which QuPID scores strictly higher. Against the strongest surrogate of Table 24, win rates are 0.58 at the full budget and 0.63 at 512, with Cliff's δ 0.19 and 0.28. At the full budget every reported interval excludes zero; at 512 examples the intervals of the 600-parameter Givens head and the tuned rotation-plane head include zero.

What the surrogates share. Every surrogate without input modulation stays in a narrow band. With 100 tuning trials, the rank-one rotation-plane head reaches 0.414 at the full budget, the Givens heads 0.409 and 0.417 at 60 and 600 parameters, and the rank-four quadratic forms 0.412 at 6,600 parameters. Neither readout rank nor a budget up to 110 times larger lifts these heads above 0.417. The circuit without re-uploading falls in the same band, reaching 0.412 with 30 parameters (Table 6). Re-uploading lifts QuPID to 0.428, above every surrogate in Table 24. Input dependence alone is not sufficient either: the spectrum-sampled features depend on the same re-uploaded projections and reach 0.407 with 328,000 parameters. The pattern points to the combination the readout is built from, input-modulated quadratic forms that share a small parameter set, rather than to rank, budget

or tuning. Table 6 uses five seeds and Table 24 uses 25 paired runs, so differences below about 0.01 across the two tables are not resolved.

Scope. The table bounds what this family of classical constructions reaches at a stated budget; it does not bound what any classical method can reach. The family contains no classical head whose rotations are modulated by the input, so it does not separate re-uploading from the circuit parameterization that implements it. A row that matches QuPID within its interval is reported as such, and the surrogates keep the shared frozen backbone, the label-free objective and the evaluation protocol of Section 4.

K Does Retrieval Quality Drive Generation

Table 25: Retrieval-utility bounds for report generation with frozen LLaVA-Med (5 seeds; std $\leq$ 0.005). *Random* draws top-5 context uniformly from the archive and *oracle* draws it from label-matched cases, which requires test-time labels and is therefore unattainable. The last column is the share of the random-to-oracle range that each retriever recovers on MIMIC-CXR CheXbert-F1. Marks and tints as in Table 1.

Retrieval context	**BLEU-4**		**CheXbert-F1**	**range**
	IU X-Ray	MIMIC-CXR	MIMIC-CXR	
no retrieval	0.112	0.081	0.246	n/a
random retrieval	0.121	0.086	0.259	0%
ViT Only	0.134	0.098	0.298	26%
Adapter	0.156	0.121	0.334	49%
QuPID	**0.163**	**0.127**	**0.352**	**61%**
oracle (label-matched)	0.198	0.161	0.411	100%

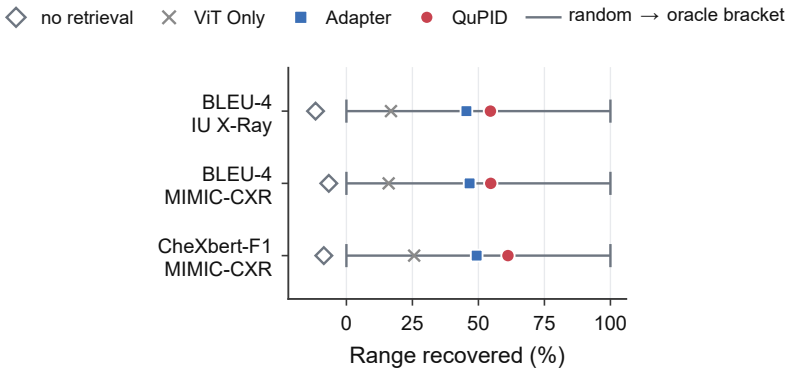


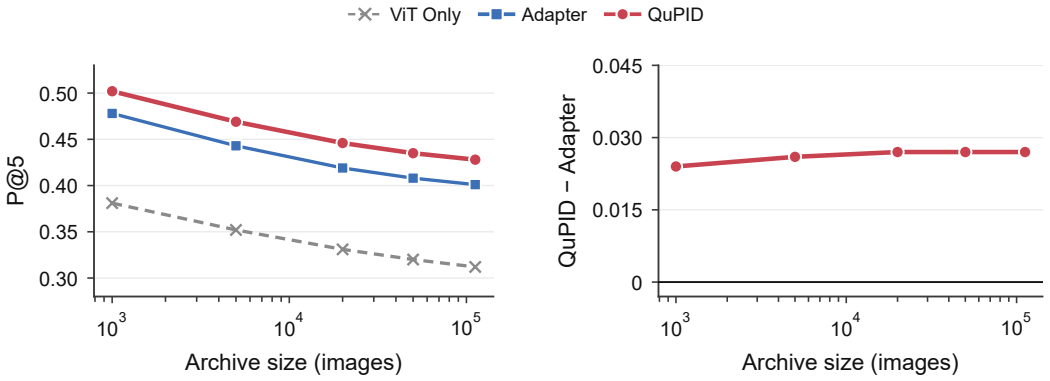
Figure 16: The retrieval-utility bracket of Table 25 for every metric the table reports (frozen LLaVA-Med, 5 seeds), normalized per metric so that random context sits at 0 and the label-matched reference at 100; the no-retrieval condition falls below random.

Section 4.1 reports generation differences of a few BLEU points between retrievers, and we first examine whether generation responds to retrieval. If random context achieved the same score as the retrieved context, the retrieval component would provide no performance benefit over random context. Table 25 brackets the effect, and Figure 16 places every retriever on the shared random-to-oracle axis. Random retrieval improves on no retrieval by only 0.009 BLEU-4, so the format of an in-context report contributes little; the label-matched oracle then adds a further 0.077, so content contributes substantially. Against that empirical reference interval QuPID recovers 61% of the clinical-proxy improvement and the classical adapter 49%. The oracle consumes test-time labels and cannot be deployed in this protocol; it provides a useful empirical reference for remaining headroom, rather than a mathematical ceiling on the performance achievable by other retrievers.

L RETRIEVAL STRESS TESTS

All retrieval results reported thus far use one archive size and one normalization scheme. This appendix varies both and examines the sequential deployment implied by data-local adaptation.

L.1 ARCHIVE SIZE



(a) P@5.

(b) Margin over Adapter.

Figure 17: Archive-size sweep (5 seeds). The horizontal axis counts stored images before patient filtering. At the 112,120-image ChestX-ray14 endpoint, the eligible gallery contains 81,664 images from 22,437 patients; query and validation patients are excluded from ranking.

A retrieval advantage measured on a small archive can disappear on a realistic one, since the number of near-duplicates that a ranking must separate grows with the gallery. Figure 17 measures retrieval as the stored cache grows. Cache size counts encoded images before split filtering, whereas ranking uses the eligible gallery after patient exclusion. The full-cache endpoint contains all 112,120 ChestX-ray14 images in storage and 81,664 images in the retrieval gallery. Absolute P@5 falls for every method, from 0.502 to 0.428 for QuPID and from 0.478 to 0.401 for the adapter, which is the expected consequence of a denser gallery. The measured margin over Adapter is $+0.024$–$+0.027$ across the cache sizes. The horizontal axis indexes storage volume, not post-filter neighbors. The same patient-exclusion rule applies throughout the sweep. Archive readouts are precomputed once per adaptation, and inference ranks eligible stored vectors. This separates readout evaluation from gallery search; the size sweep reports accuracy, not a constant-latency guarantee.

Cache accounting and patient exclusion. The complete ChestX-ray14 cache partitions into 81,664 gallery, 16,384 query and 14,072 validation images. The corresponding disjoint patient counts are 22,437, 4,504 and 3,864. The 32,768 adaptation images belong to the gallery and are not another disjoint partition. Cache construction and split membership are separate: query and validation images may be encoded for evaluation, but their patient IDs are excluded from the neighbor index. Smaller points in Figure 17 retain their raw-cache axis values and apply the same exclusion rule; their eligible counts need not equal those axis values.

L.2 FEATURE NORMALIZATION AND NORM SHIFT

Amplitude encoding requires strict L2 normalization, raising the concern that the method inherits a sensitivity to feature scale that classical modules avoid. Table 26 varies the preprocessing and reports a P@5 range of at most 0.012 across the preprocessing choices. Whitening is marginally better than the default (0.431 against 0.428), while PCA-1024 is marginally worse (0.426) and mean removal is worst (0.419). Whitening changes both basis and scale; full-dimensional orthogonal PCA need not discard information. The measured differences therefore probe the structured head's sensitivity to preprocessing, rather than establishing reconstruction loss as their cause. We keep plain L2 as the default rather than whitening, since the 0.003 difference is within the observed variation across seeds and whitening needs a second pass over the archive to estimate the covariance, which adds preprocessing cost even when performed locally. Table 27 evaluates sensitivity to feature gain and offset shifts. A global rescaling is annihilated by normalization by construction and would not

Table 26: Encoding variants at the full budget (P@5, ChestX-ray14, 5 seeds; std $\leq$ 0.007). Default: L2 normalization only. Marks and tints as in Table 1.

Preprocessing	P@5
L2 only (default)	0.428
whitening, then L2	**0.431**
PCA-1024, then L2	0.426
mean removal, then L2	0.419

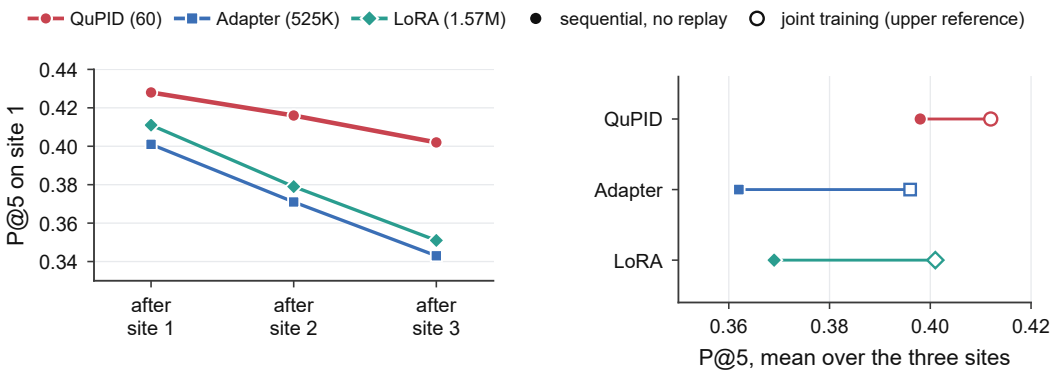
Table 27: Gain and offset shift applied to query features before normalization (P@5, 5 seeds; std $\leq$ 0.008). Per-dimension gains are drawn around s with a constant offset, so the shift is not annihilated by L2 normalization. Marks and tints as in Table 1.

Method	s=0.80	0.90	1.00	1.10	1.25
Adapter	0.392	0.398	0.401	0.397	0.389
QuPID	**0.421**	**0.426**	**0.428**	**0.425**	**0.418**
Δ	+0.029	+0.028	+0.027	+0.028	+0.029

Normalization removes a common positive scale, but it does not remove an additive offset applied before normalization. The paired tables distinguish that invariance from changes that rotate the normalized feature direction.

test sensitivity to feature-scale shifts, so we instead draw per-dimension gains around s and add a constant offset, which normalization cannot undo. Both methods lose accuracy at the extremes and the margin stays within 0.002 of its unshifted value. This covers gain and offset shift only. Figure 5(c) takes up distributional shift of the features.

L.3 Sequential site adaptation



(a) Site 1 after each stage.

(b) Mean over sites against joint training.

Figure 18: Sequential adaptation over three sites without replay, re-evaluating site 1 after each stage (5 seeds); filled markers in (b) are the final module averaged over the three sites and hollow markers joint training on the pooled data.

Table 28: Sequential adaptation over three sites (ChestX-ray14, CheXpert, MIMIC-CXR) with no replay buffer and no continual-learning method (P@5, 5 seeds; std $\leq$ 0.009). The first three columns evaluate on site 1 after each stage; *mean* averages the final module over all three sites; *joint* trains one module on the pooled data as an upper reference. Marks and tints as in Table 1.

Method	p	site 1	+site 2	+site 3	forget	mean	joint
Adapter	525K	0.401	0.371	0.343	0.058	0.362	0.396
LoRA (r=16)	1.57M	0.411	0.379	0.351	0.060	0.369	0.401
QuPID	60	**0.428**	**0.416**	**0.402**	**0.026**	**0.398**	**0.412**

Data-local adaptation implies that sites arrive one after another and that data from an earlier site cannot be revisited, which is the setting in which a stored module must remain effective. Table 28 and Figure 18 adapt each module on three sites in sequence with no replay, and re-evaluate site 1 after each stage. The readout loses 0.026 P@5 on the first site by the end against 0.058 and 0.060 for the two adapters, and its final module is the best of the three on the average over all sites. This is consistent with a capacity-based explanation, but parameter count alone cannot guarantee less forgetting: even a small module can change its function substantially. Two scope statements belong with this result. We apply no continual-learning method, no replay, and no regularization toward previous solutions, so this measures the raw budget effect and not a competitive continual-learning system. Joint training on pooled data remains better than the sequential result for every method, so this compares deployments that must forget rather than arguing for sequential adaptation.

M Compute Parity and Statistical Protocol

M.1 Equal wall-clock instead of equal parameters

Table 29: Classical baselines given QuPID's adaptation wall-clock instead of its parameter count (P@5, ChestX-ray14, 5 seeds; std ≤ 0.008). Extra epochs and random restarts are selected by contrastive validation loss. QuPID is unchanged. Marks and tints follow Table 1.

Method	n=512		**full budget**		**epochs**
	default	matched	default	matched	
Adapter	0.343	0.351	0.401	0.404	50 → 80
LoRA (r=16)	0.347	0.354	0.411	0.414	50 → 68
Adapter-XL	0.338	0.349	0.414	0.417	50 → 61
QuPID	**0.401**		**0.428**		50
QuPID margin	+0.054	+0.047	+0.014	+0.011	

Extra compute improves each included classical baseline while preserving the measured ordering. The separate rotation-plane comparison below evaluates the stronger 60-parameter control with validation-selected tuning and its own trial budget.

Table 9 states that parameter efficiency is not computational efficiency, and Table 29 addresses the corresponding compute-budget comparison. Because statevector simulation makes an adaptation epoch roughly 1.6 times more expensive, a comparison at equal parameter counts allocates less computation time to the baselines. The three listed classical modules receive the adaptation wall-clock of QuPID, spent on extra epochs and random restarts selected by contrastive validation loss. The rotation-plane comparison below specifies its own trial budget and retains the same fixed QuPID reference. The baselines gain 0.003 to 0.011 P@5. Against the three modules in this table, our margin at $n = 512$ falls from $+0.054$ to $+0.047$ and at the full budget from $+0.014$ to $+0.011$; these margins exclude the controls in Table 23. The ordering does not change. These results distinguish parameter count, the duration of an individual adaptation run and the cost of searching multiple configurations. Their margins apply to the methods and selection budgets stated for each comparison.

Rotation-plane tuning budget. The rotation-plane tuning study allows 30 contrastive-validation trials, with a per-trial adaptation cap of 60 seconds at $n = 512$ and 3,600 seconds at full budget, including restarts. The fixed QuPID configuration is retained as the reference; its reported full-budget training time is $50 \times 71.3 = 3{,}565$ seconds. These are per-trial ceilings, not a claim of identical total search time. Both methods use the same frozen-feature inputs and validation criterion; cached-feature extraction is outside the adaptation timer.

n	Rotation default	Rotation tuned	QuPID fixed
512	.386	.394	.401
32,768	.405	.414	.428

The tuned rotation-plane scores are 0.394 at 512 and 0.414 at full budget, leaving margins $+0.007$ and $+0.014$ against fixed QuPID. Table 24 reports paired 95% intervals for this tuned configuration, $[-0.004, +0.018]$ at 512 and $[+0.006, +0.022]$ at full budget, so the tuned margin excludes zero at the full budget and not at 512. The intervals below concern the default rotation-plane scores 0.386 and 0.405 and are not transferred to the tuned comparison.

M.2 Variance decomposition and effect sizes

Table 30: Variance components of P@5 from a crossed design of 5 initializations by 5 adaptation-subset draws ($\times 10^{-4}$, QuPID on ChestX-ray14).

n	**subset**	**init.**	**inter.**	**subset share**
512	6.8	1.9	0.7	72%
32,768	0.9	1.1	0.3	39%

Table 31: Effect sizes for per-query paired scores, alongside the significance test in Section 4. Win rate is the share of queries on which QuPID scores strictly higher.

Comparison	n	**Cliff's δ**	**win rate**
vs. sweep baseline	512	0.44	0.67
vs. Adapter-XL	32,768	0.21	0.58

A seed in our protocol controls initialization, augmentation, and the adaptation-subset draw at once, which is suitable for reporting aggregate variability but does not separate its sources. Table 30 separates the two factors with a crossed design of five initializations by five subset draws. At $n = 512$ the subset draw accounts for 72% of the variance and initialization for most of the remainder; at the full budget the two are comparable. Query bootstrapping conditional on five fitted runs omits variation from drawing a new adaptation set. The paired analysis uses repeated subset and initialization draws for QuPID and the default rotation-plane head, using held-out patients as clusters.

Table 31 reports effect sizes next to the paired bootstrap of Section 4, since Holm correction controls family-wise type-I error, while rejection alone does not quantify the size of a difference. Cliff's δ is 0.44 at $n = 512$, a medium effect, and 0.21 at the full budget, a small one; the per-query win rates are 0.67 and 0.58. Both orderings are consistent with the stated claim, and the statistically significant full-budget difference remains small: a win rate of 0.58 leaves roughly two queries in five as losses or ties, so the average gain is not a per-query guarantee. The two statistics answer different questions and we report both deliberately. The bootstrap and effect sizes are conditional summaries for the included comparators. They neither describe new adaptation draws nor establish a paired advantage over the rotation-plane head. Multiple queries from one patient further invalidate an independent-query interpretation of population uncertainty. A third perspective concerns the deployment setting. Since the comparison is paired per query, the practically relevant quantity is the gain conditional on the adapter failing, and there QuPID recovers a relevant case in the top five on 0.31 of the queries where the adapter retrieves none at $n = 512$, against 0.19 in the reverse direction.

Paired estimand and evaluation units. The estimand is the mean P@5 difference between QuPID and the default rotation-plane head over adaptation draws, initialization and held-out patients. Each method is evaluated on the same adaptation subset and query manifest. The point estimate averages queries equally within each fitted run and then averages the crossed runs; patient IDs define clusters for the uncertainty calculation. The design crosses five adaptation-subset draws with five initialization draws, giving 25 paired fitted-run comparisons per budget. Pairing retains one score per query, method, subset draw and initialization. The supplementary resampling implementation supports nested and crossed designs. For nested initialization it resamples subset IDs followed by paired initialization replicates; for crossed initialization it resamples both factor levels independently (Owen, 2007). Crossed cells are not treated as independent training replicates. Patient resampling retains all images of a sampled patient and uses the same draw for both methods and all fitted runs. This preserves within-patient dependence while maintaining query-weighted point estimates. The numerical intervals reported below use a normal approximation to the measured variance components. The supplementary code also implements percentile resampling; the displayed intervals are not percentile-bootstrap output. The analysis conditions on the gallery, validation split and default hyperparameters; variability from the separate tuning search is outside its scope.

Normal-approximation intervals. The analysis uses the 5×5 crossed runs and $4{,}504$ held-out patient clusters. At the reported precision, the rotation-plane variance components equal those of QuPID, a common between-method correlation estimate $\rho = 0.5$ is used for the training components, and the patient-level contribution to the difference variance is $v_p = 0.01$. With subset, initialization and interaction components v_s, v_i, v_{si} from Table 30, the estimated variance of the mean difference is $2(1 - \rho)(v_s/5 + v_i/5 + v_{si}/25) + v_p/4{,}504$. The intervals use the mean difference plus or minus 1.96 standard errors.

n	Mean margin	SE	Normal 95% interval
512	+0.015	0.0134	$[-0.0112, +0.0412]$
32,768	+0.023	0.0066	$[+0.0101, +0.0359]$

At $n = 512$, the interval includes zero despite the positive mean margin, so this analysis does not establish a positive difference at 95%. The full-budget interval is positive. Both intervals summarize the default paired comparison under the stated approximation and fixed evaluation split.

N Design Choices and Empirical Scope

The circuit construction, training protocol and deployment path determine the scope of the reported evidence. Tables 32 and 33 connect each design choice and empirical comparison to its supporting result. The central contribution is a structured feature map that combines a small train-

able budget with competitive retrieval quality. Larger adapters assess the accuracy attainable at a much greater budget; equally compact classical heads assess the chosen parameterization. All comparisons use classical simulation and do not establish computational quantum advantage. The rotation-plane sweep measures accuracy across adaptation sizes, its separate tuning study reports validation-selected scores, and the paired intervals quantify uncertainty for the default endpoints. The full-budget interval is positive, while the 512 interval includes zero. Split counts distinguish stored cache size from the eligible gallery.

Table 32: Design choices, empirical interpretation and supporting results (I: model, controls, trainability, cost, deployment, encoding and statistics). References point to sections and appendices in this paper; Table 33 continues the list.

Aspect	Interpretation	Evidence
Circuit-defined model class	The circuit shares 60 parameters across $M = 40$ input-modulated, correlated full-rank quadratic forms. Its contribution is this compact retrieval feature map, supported by competitive ranking quality under classical simulation.	Sec. 1, 3.3, 4.2; App. A.6
Classical control budgets	The rotation-plane head matches 60 parameters, isometry and output width. Its input-independent rank-one forms differ from the input-modulated full-rank Pauli forms. The comparison evaluates these choices jointly; RFF and the four-projection Quadratic head are more restricted controls.	Table 1; App. J.2, J.3, M.1
Contribution beyond frozen features	The backbone initialization is shared; post-encoder heads keep features frozen, while LoRA updates attention projections; the untrained circuit and the naive fidelity design are reported as controls, and the naive design exactly matches the frozen baseline as the theory requires; all 60 parameters are updated.	Fig. 4(a); Table 6; App. B.4, G
Finite-scale trainability	Depth $L \leq 6$, $n_q = 10$ fixed by the feature dimension, local one- and two-body observables; gradient variance is a finite diagnostic, not an asymptotic rate. Classical simulation differentiates the statevector; parameter shift provides the gate-level hardware cost model.	Sec. 3.4; Fig. 11; Table 13
Simulation cost and scaling	The statevector dimension is $2^{n_q} = d$, with runtime also scaling in depth, gate count and readout width; per-epoch training time is $1.60\times$ that of a light adapter, inference time and memory usage are comparable, and the qubit count is swept rather than assumed.	Table 9; Fig. 12(a); App. H.1
GPU deployment and noise simulation	The intended system trains by classical circuit simulation and runs fixed-parameter GPU inference with cached archive readouts. No quantum processor is required, as in quantum parameter-efficient fine-tuning (Koike-Akino et al., 2025; Liu et al., 2025). Noise and shot studies are simulations of an optional hardware realization, not its validation.	Sec. 2, 4.2; Fig. 5(a,b); App. A.6, I
Amplitude encoding in simulation	In simulation it is a normalization; its real costs, strict L2 normalization and power-of-two dimensions, are stated, with the padding a 768-dimensional backbone would need.	Sec. 5; App. A.2, L.2
Effect sizes and uncertainty	Query-bootstrap markers condition on fitted runs. Normal intervals against the default rotation-plane head include adaptation, initialization and patient variation: positive at full budget, inconclusive at 512. No paired interval is reported for the tuned control.	Fig. 4(b), 13; App. M.2

Table 33: Design choices, empirical interpretation and supporting results (II: theory, surrogates, circuit design, shots, datasets, low-data accuracy, generation and scope).

Aspect	Interpretation	Evidence
Scope of the generalization bound	The bound applies to bounded smooth parameter classes and concerns contrastive risk. It does not certify the reported runs or predict P@5 gaps. The empirical claim combines a small budget with competitive retrieval; the gap ratios are separate diagnostics.	Sec. 3.4; App. B.3, J.1
Classical surrogate comparisons	The default rotation-plane head leaves full/512-example margins of $+0.023/+0.015$; tuning reduces them to $+0.014/+0.007$. Appendix J.4 extends this to surrogates with far larger budgets and more tuning; without input modulation they stay between 0.409 and 0.417 P@5 up to $6{,}600$ parameters. These measured comparisons do not establish a classical expressivity separation (App. M.1).	Sec. 4.2; Table 1, 24
Circuit and readout selection	Real rotations keep the state real and the simulation computationally inexpensive; qubit count, entangling topology, readout width and depth are each swept and saturate at the default, so the operating point is a plateau.	Fig. 4(a), 12; App. A.3, H.1
Finite-shot readout	A sufficient Hoeffding ranking guarantee is derived with explicit norm-floor, margin and collection-size dependence. Separately, the finite-shot experiment is within 0.002 P@5 of exact expectations at 10^4 shots per observable.	Sec. 4.2; App. I
Dataset coverage	ChestX-ray14 (112,120 images) and MURA for retrieval, IU X-Ray and MIMIC-CXR for generation, MVTec AD and CUB-200 for generality.	Sec. 4.1; App. C
Low-data accuracy and degradation	The sweep tests retrieval quality as adaptation data shrink. Default rotation-plane margins are $+0.023$ at full budget and $+0.015$ at 512. The control degrades less; the results support observed accuracy at a small budget, without establishing superior stability.	Sec. 3.4, 4.1; App. J.1, J.2
Retrieval-grounded generation	The pipeline is fixed and stated: cached archive readouts, top-k reports concatenated with the query image, a frozen generator with greedy decoding. Random and label-oracle contexts provide an empirical reference range. Clinical-efficacy proxies assess factual content, and per-query costs are reported for both methods.	Sec. 4.1; Table 8, 9; App. K
Interpretation boundaries	Accuracy claims concern the stated ranking metrics and comparison budgets. Raw cache size differs from the patient-filtered gallery. Data-local GPU execution is not a formal privacy guarantee.	Sec. 4.1, 5
Relation to quantum adaptation	Quantum-PEFT and quantum parameter adaptation act on the weight side and never encode data; projected kernels measure a fixed encoding. Here data are encoded and re-uploaded, and the circuit learns a retrieval feature map; Proposition 1 shows why measurement escapes shared-unitary overlap invariance.	Sec. 2; App. A.6
Background and implementation	A self-contained background appendix introduces the ingredients in method order and provides a runnable listing.	App. A