TRUST-AWARE PROTOTYPE-GUIDED QUANTUM REPRESENTATION LEARNING FOR BACKDOOR-RESISTANT MEDICAL IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-time backdoors turn small image artifacts into prediction shortcuts, and compact medical classifiers are especially exposed because a low-dimensional representation leaves little room to absorb a spurious direction. We present **TAPQ-Net**, a hybrid quantum–classical defense that trains directly on poisoned data, without a trusted clean subset, sample removal, or post-hoc purification. A frozen encoder and a fixed PCA projection map each image to eight coordinates. A label-free expectation–maximization (EM) mixture assigns every sample a soft prototype and a trust score; an 8-qubit tree-structured quantum encoder (TQE) with seven trainable rotations is aligned to that prototype, and the trust score bounds the sample's cross-entropy weight. The mechanism that separates TAPQ-Net from prior prototype- and clustering-based defenses is a reverse channel: the EM state may evolve only when the silhouette geometry of the quantum embedding stays separated, balanced, and non-collapsed. Because EM never reads labels, relabeling reaches the prototypes only through the trigger's latent offset, and alignment bounds the trigger-direction component of the embedding. On OASIS and six MedMNISTv2 datasets under random-class and targeted patch poisoning at 1–10%, TAPQ-Net attains the lowest attack success rate in all 42 conditions (0.13–0.24, against at least 0.34 for every compared method) and the highest clean accuracy in 40, and paired tests over the 42 conditions are significant for all four metrics. A topology- and parameter-matched classical tree trails the TQE by 3.4 accuracy points, poisoned samples receive lower trust without being removed, and the defense carries over to clean-label and warping-based triggers.

1 INTRODUCTION

The reliability of medical image classifiers depends on the integrity of their training data as much as on model capacity (Litjens et al., 2017; Topol, 2019; Karimi et al., 2020). Medical datasets are small, noisily labeled, class-imbalanced, and separated by subtle visual differences (Gao et al., 2020; Rajaraman et al., 2022; Ju et al., 2022), so a few corrupted samples can distort the decision boundary (Finlayson et al., 2019; Dong et al., 2024). Training-time backdoor poisoning exploits exactly this: trigger-stamped, relabeled training images yield a model that behaves normally on clean inputs and obeys the attacker whenever the trigger reappears (Gu et al., 2017; Wu et al., 2022), even under label-only or multimodal poisoning (Jha et al., 2023; Yang et al., 2023b), and multi-site aggregation, multi-reader annotation, and reused pretrained models make this threat concrete in medicine (Rieke et al., 2020; Irvin et al., 2019). Existing defenses act on samples, inputs, or a repaired model (Sec. 2) and leave open a representation-level question that is acute for compact medical models: *how can a low-dimensional embedding preserve disease-relevant structure while preventing a small trigger from becoming a shortcut feature?*

Hybrid quantum–classical models make this question concrete: a classical encoder compresses the image into a few latent coordinates, and a variational quantum circuit (VQC) supplies a low-parameter nonlinear embedding (Benedetti et al., 2019; Landman et al., 2022; Bokhan et al., 2022; Dhara et al., 2024; Ajlouni et al., 2023; Yousif et al., 2024; Rahman & Zhuang, 2026a). We adopt a tree-structured quantum encoder (TQE), whose pair–branch–root aggregation distills local information into global structure (Grant et al., 2018; Huggins et al., 2019; Cong et al., 2019). Compactness is also the

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

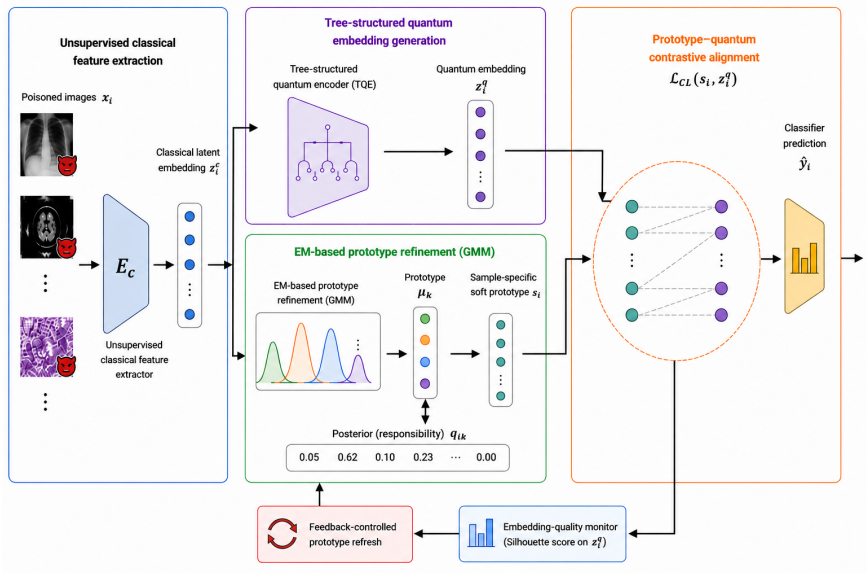


Figure 1: **Overview of TAPQ-Net.** *Unsupervised classical feature extraction*: the frozen extractor E_c and a fixed PCA projection map each poisoned image x_i to the classical latent embedding z_i^c. *Tree-structured quantum embedding generation*: the TQE maps z_i^c to the quantum embedding z_i^q (Eq. (4)). *EM-based prototype refinement*: a label-free Gaussian mixture over $\{z_i^c\}$ yields prototypes μ_k, responsibilities q_{ik}, and a sample-specific soft prototype s_i with trust score r_i (Eqs. (1)–(3)). *Prototype–quantum contrastive alignment*: $\mathcal{L}_{\mathrm{CL}}(s_i, z_i^q)$ pulls z_i^q toward s_i, and the classifier predicts $\widehat{y}_i$ under r_i-weighted cross-entropy (Eq. (5)). *Embedding-quality monitor* and *feedback-controlled prototype refresh*: after each epoch, the silhouette score of $\{z_i^q\}$, together with component balance and activity, decides whether the EM state is refreshed or retained (Eq. (6)).

weakness: with eight coordinates, a trigger-correlated direction competes directly with disease-relevant structure, and hybrid models are demonstrably backdoor-susceptible (Guo et al., 2025a). Robust hybrid learning must therefore regulate the *geometry* of the quantum embedding; prototypes anchor embeddings to semantic centers in low-data regimes (Snell et al., 2017; Li et al., 2021a; Jin et al., 2022), but prototypes estimated from poisoned data absorb trigger-correlated samples, and in existing defenses the geometry that judges samples is fixed once estimated or recomputed unconditionally: none lets the representation being trained decide whether its own supervisory targets may change.

We propose **TAPQ-Net**, which closes this loop through three mechanisms (Fig. 1). **First**, a label-free EM procedure refines soft prototypes on the classical latent vectors, estimating posterior responsibilities and a per-sample trust score. **Second**, an 8-qubit TQE with seven trainable rotations generates a constrained, hierarchically aggregated quantum embedding. **Third, and crucially**, prototype–quantum contrastive alignment pulls each quantum embedding toward its soft prototype, and an embedding-quality monitor on the quantum geometry drives a feedback-controlled prototype refresh that decides whether the prototype state may take its next EM step. Our contributions are:

- **A failure mode and its remedy.** We identify how poisoning corrupts both the prototype targets and the quantum embedding of compact hybrid models, and introduce feedback-controlled prototype refresh, in which the embedding's separation, balance, and non-collapse decide whether its targets may evolve (Sec. 3).
- **Trust-aware soft prototypes with formal bounds.** We show that label-free EM confines the attacker's influence on the prototypes to the trigger's latent offset and that alignment bounds the trigger-direction component of the embedding (Sec. 4, App. A).
- **A structured quantum encoder under prototype supervision.** We show that the TQE is not interchangeable: a parameter-matched MLP, a flat VQC, and a classical tree with identical topology and budget all fall short of it inside the same loop (Sec. 5).
- **Evidence across 42 conditions.** Against eleven baselines, TAPQ-Net attains the lowest ASR in all 42 conditions and the highest accuracy in 40, with significant paired gains; trust diagnostics, a PCA sweep, and clean-label and warping attacks delimit the gains (Sec. 5).

2 Related Work

Backdoor defenses and latent geometry. Backdoor attacks implant trigger-dependent behavior while preserving clean accuracy (Gu et al., 2017; Wu et al., 2022). Detection-and-filtering methods cluster activations or frozen features and retrain on the retained set (Tran et al., 2018; Chen et al., 2019; Kutbi, 2025); post-hoc defenses repair a backdoored model or its inputs (Zhu et al., 2023; Wei et al., 2023; Lin et al., 2024; Amula et al., 2025; Shi et al., 2023; Yang et al., 2024), although low ASR after purification need not remove backdoor-connected features (Min et al., 2024). In-training defenses isolate, decouple, or reweight samples (Li et al., 2021b; Huang et al., 2022b; Zhang et al., 2023), and PGRL derives prototypes from a verified-benign subset (Guo et al., 2025b). TAPQ-Net shares the frozen-feature, PCA, and Gaussian-mixture front end of Kutbi (2025) but removes nothing, needs no clean subset or auxiliary backdoored model, feeds responsibilities into training as continuous targets and weights, and re-estimates the mixture only when the representation it supervises passes a geometric gate (App. E).

Quantum robustness and quantum medical models. Quantum robustness research addresses adversarial examples through noise and certification (Du et al., 2021; Huang et al., 2022a; Sahdev & Kumar, 2023) and poisoning through detection and attack analysis (He et al., 2025; Kundu & Ghosh, 2025); ACDM and QSentry detect QNN backdoors by clustering measurement outputs (Zhang, 2026; Wang et al., 2025). Among medical quantum models (Mathur et al., 2021; Landman et al., 2022; Bokhan et al., 2022; Dhara et al., 2024; Ajlouni et al., 2023; Yousif et al., 2024), NQNN and QAPNet adapt quantum branches to corrupted supervision and corrupted inputs, with QAPNet using prototype-contrastive regularization (Rahman & Zhuang, 2026a;b). None addresses backdoor poisoning with feedback-controlled prototype supervision.

Prototypes and hierarchical circuits. Prototypes and EM-style contrastive learning provide low-data inductive bias and semantic compactness (Snell et al., 2017; Li et al., 2021a; Jin et al., 2022); tree tensor networks and hierarchical quantum classifiers aggregate local information progressively and ease trainability (Orús, 2014; Stoudenmire & Schwab, 2016; Grant et al., 2018; Huggins et al., 2019; Cong et al., 2019; McClean et al., 2018; Cerezo et al., 2021; Pesah et al., 2021) (App. I).

3 Method

Fig. 1 summarizes TAPQ-Net; the paragraphs below follow its blocks, and Sec. 4 analyzes why the resulting loop resists the trigger.

Threat model. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with $y_i \in \mathcal{Y} = \{1, \dots, C\}$. The attacker selects an unknown index set $\mathcal{P}$ with $|\mathcal{P}| = \eta N$, replaces x_i by $T(x_i)$ for an image-space trigger T, and relabels each poisoned sample with a label drawn from $\mathcal{Y} \setminus \{y_i\}$ (random-class) or with a fixed target t (Targeted). The defender observes the poisoned set $\mathcal{D}^p$, holds no trusted clean subset, and knows neither $\mathcal{P}$, T, η, nor the mapping. A clean validation split serves only for epoch selection.

Unsupervised classical feature extraction. The classical feature extractor E_c is a frozen pretrained ResNet-18 that maps each image to $f_i = E_c(x_i) \in \mathbb{R}^{512}$. We call this stage unsupervised because neither E_c nor the projection that follows ever reads a training label. We fit a PCA projection $W \in \mathbb{R}^{8\times512}$ once on the features of the poisoned training split and then freeze it, giving $z_i^c = W(f_i - \bar{f}) \in \mathbb{R}^{d_z}$ with $d_z = 8$; validation and test data never enter the fit. PCA thus compresses the 512-dimensional pooled features, not image pixels: for a 224×224 OASIS slice, the frozen encoder performs the pixel-to-feature reduction. On PathMNIST the eight components retain 93.2% of the feature variance, and fewer components lower accuracy monotonically (Tab. 4b); $d_z = 8$ also permits one-to-one angle encoding on eight qubits and a readout matching the prototype dimension.

EM-based prototype refinement. We model $\{z_i^c\}$ with a Gaussian mixture of $K = C$ isotropic components $\{\pi_k, \mu_k, v_k\}_{k=1}^{K}$, initialized by clustering $\{z_i^c\}$ into K groups. With a temperature $\tau > 0$ that softens the responsibilities, the E-step computes

$$\ell_{ik} = \frac{1}{\tau}\Big[-\frac{\|z_i^c - \mu_k\|_2^2}{2v_k} - \frac{d_z}{2}\log(2\pi v_k) + \log \pi_k \Big], \qquad q_{ik} = \frac{\exp(\ell_{ik})}{\sum_{m=1}^{K}\exp(\ell_{im})}, \tag{1}$$

and, with $M_k = \sum_i q_{ik}$, the M-step computes

$$\widehat{\pi}_k = \frac{M_k}{\sum_m M_m}, \qquad \widehat{\mu}_k = \frac{1}{M_k}\sum_i q_{ik} z_i^c, \qquad \widehat{v}_k = \frac{1}{M_k d_z}\sum_i q_{ik}\|z_i^c - \widehat{\mu}_k\|_2^2. \tag{2}$$

Each EM pass updates $\pi_k \leftarrow (1-\alpha)\pi_k + \alpha\widehat{\pi}_k$ and applies the same exponential moving average (EMA) to μ_k and v_k for components with $M_k \geq m_{\min}$. From the accepted state we form the trust score r_i, the monitoring assignment a_i, and the sample-specific soft prototype s_i:

$$r_i = \max_k q_{ik}, \qquad a_i = \arg\max_k q_{ik}, \qquad \bar{s}_i = \sum_k q_{ik}\mu_k, \qquad s_i = r_i\bar{s}_i + (1-r_i)z_i^c. \tag{3}$$

A decisive assignment pulls the target toward the mixture mean; an ambiguous one keeps it near the observation. Labels enter neither Eq. (1) nor Eq. (2), and a_i serves only for monitoring. Hence r_i measures assignment concentration under the current mixture geometry and is *not* a probability of being clean.

Tree-structured quantum embedding generation. The TQE prepares $|\psi_i^{(0)}\rangle = \bigotimes_{j=0}^{7} R_Y(z_{i,j}^c)|0\rangle$ and applies three interaction levels $\mathcal{P}_1 = \{(Q_0,Q_1),(Q_2,Q_3),(Q_4,Q_5),(Q_6,Q_7)\}$, $\mathcal{P}_2 = \{(Q_1,Q_3),(Q_5,Q_7)\}$, and $\mathcal{P}_3 = \{(Q_3,Q_7)\}$; each pair (u,w) acts through the block $B_{(u,w)}^{(\ell)} = R_Y(\theta_w^{(\ell)})\,\mathrm{CNOT}_{u\to w}$ (Fig. 2). The readout and classifier are

$$z_i^q = \big[\langle Z_0\rangle_i, \ldots, \langle Z_7\rangle_i\big]^\top \in [-1,1]^8, \qquad \widehat{y}_i = \mathrm{softmax}\big(W_{\mathrm{cls}}\, z_i^q + b\big). \tag{4}$$

The circuit has seven trainable angles (Tab. 8, App. C.3). Because a CNOT commutes with Z on its control, the leaf controls Q_0, Q_2, Q_4, Q_6 retain $\langle Z_j\rangle_i = \cos z_{i,j}^c$, and learned information concentrates along the target path—one rotation on Q_1 and Q_5, two on Q_3, three on the root Q_7 (Remark A.2). The embedding thus combines a fixed cosine feature map with a learned local-to-global aggregation; each PCA coordinate enters directly as a rotation angle.

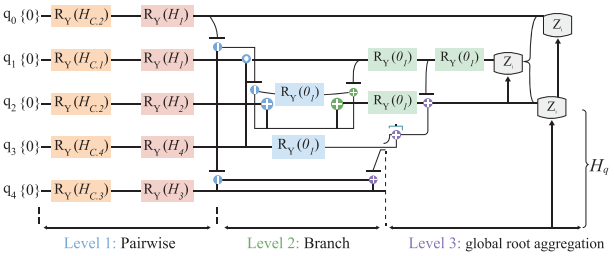


Figure 2: **Tree-structured quantum encoder.** z_i^c is angle-encoded on eight qubits, aggregated at pairwise ($\mathcal{P}_1$), branch ($\mathcal{P}_2$), and root ($\mathcal{P}_3$) levels, and read out as Pauli-Z expectations z_i^q.

Prototype–quantum contrastive alignment. The alignment block pulls every quantum embedding toward its own soft prototype and trains the classifier under trust-weighted supervision. For a mini-batch $\mathcal{B}$, with the prototype state held fixed so that no gradient reaches s_i or r_i, we minimize

$$\mathcal{L}_{\mathrm{CL}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\Big(1 - \frac{(z_i^q)^\top s_i}{\|z_i^q\|_2\,\|s_i\|_2}\Big), \quad \mathcal{L}_{\mathrm{CE}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\bar{r}_i\,\ell_{\mathrm{CE}}(\widehat{y}_i, y_i), \quad \mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{CL}}\mathcal{L}_{\mathrm{CL}}, \tag{5}$$

where $\bar{r}_i = \max(r_i, r_{\min})$, the floor $r_{\min} > 0$ keeps every sample in the objective, and $\lambda_{\mathrm{CL}} > 0$ weights the alignment. $\mathcal{L}_{\mathrm{CL}}$ contrasts each embedding with its own sample-specific target rather than a shared centroid and is invariant to the scale of z_i^q. Gradients update only θ and (W_{cls}, b).

Embedding-quality monitor and feedback-controlled prototype refresh. After every epoch t, the embedding-quality monitor passes the full poisoned training set through the TQE and computes three label-free statistics from $\{z_i^q\}$ and the monitoring assignments $\{a_i\}$: the silhouette coefficient $SS_t = \mathrm{Silhouette}(\{z_i^q\}, \{a_i\})$, the largest-component ratio $\rho_t = \max_k m_k/N$ with $m_k = |\{i : a_i = k\}|$, and the number of active components $K_{\mathrm{act}} = |\{k : m_k \geq m_{\min}\}|$. With $SS_{\mathrm{best}}^{(t-1)} = \max_{u<t} SS_u$, the feedback-controlled refresh executes one further EM pass only if

$$t \bmod \Delta_{\mathrm{ref}} = 0, \qquad SS_t \geq SS_{\mathrm{best}}^{(t-1)} - \varepsilon_{\mathrm{sil}}, \qquad \rho_t \leq \rho_{\max}, \qquad K_{\mathrm{act}} \geq K_{\min}, \tag{6}$$

and otherwise retains the previous state. Because $\{z_i^c\}$ is fixed, a refresh is one further EMA-damped EM pass on the same data: the gate decides whether the mixture may continue to evolve, and it decides so from the geometry of the representation that the mixture supervises. Neither labels nor validation data enter the gate. Every epoch thus trains the TQE and classifier with the EM state frozen, recomputes $\{z_i^q\}$, and runs one EM pass only if Eq. (6) holds; inference uses only E_c, PCA, the TQE, and the classifier (Algorithms 1–2, Tab. 6).

4 THEORETICAL ANALYSIS

We analyse one accepted EM state; all proofs are in App. A. For analysis only, let $c(i)$ be the semantic class of sample i, let $\kappa(c) \in \{1, \dots, K\}$ be the component matched to class c by a post-hoc one-to-one matching, and, for $\kappa = \kappa(c)$, let $w_i = q_{i\kappa}/M_\kappa$, so that $\sum_i w_i = 1$ and $\widehat{\mu}_\kappa = \sum_i w_i z_i^c$ (Eq. (2)); the EM procedure itself never uses labels. Write $\mathcal{C}_c = \{i \notin \mathcal{P} : c(i) = c\}$, $\mathcal{P}_c = \{i \in \mathcal{P} : c(i) = c\}$, $\mathcal{O}_c = \{i : c(i) \neq c\}$, and $n_c = |\mathcal{C}_c \cup \mathcal{P}_c|$.

Assumptions. **(A1)** *Latent model*: $z_i^c = \nu_{c(i)} + \xi_i + \mathbf{1}[i \in \mathcal{P}]\,\delta_i$, where the ξ_i are independent, zero-mean, and ς-sub-Gaussian, and the trigger offset $\delta_i = z^c(T(x_i)) - z^c(x_i)$ satisfies $\|\delta_i\|_2 \leq \beta$. **(A2)** *Contamination*: $\sum_{i \in \mathcal{P}_c} w_i \leq \eta_\kappa$. **(A3)** *Leakage*: $\sum_{i \in \mathcal{O}_c} w_i \leq \gamma$, and $\|z_i^c - \nu_c\|_2 \leq \Delta$ for every $i \in \mathcal{O}_c$. **(A4)** *Weight regularity*: the weights are fixed with respect to $\{\xi_i\}_{i \in \mathcal{C}_c \cup \mathcal{P}_c}$, and $\sum_{i \in \mathcal{C}_c \cup \mathcal{P}_c} w_i^2 \leq c_0/n_c$. (A2) holds with $\eta_\kappa \approx \eta$ when poisoned and clean samples of a class receive comparable responsibilities, and (A4) idealizes the dependence of the responsibilities on the data.

Proposition 4.1 (Prototype deviation). *Under (A1)–(A4), with probability at least $1 - \delta$, simultaneously for all classes c,*

$$\left\|\widehat{\mu}_{\kappa(c)} - \nu_c\right\|_2 \;\leq\; C_1\, \varsigma \sqrt{\frac{c_0\,(d_z + \log(K/\delta))}{n_c}} \;+\; \eta_\kappa \beta \;+\; \gamma\Delta,$$

where C_1 is a universal constant. If the initial means satisfy the same bound, it holds for the EMA-stabilized means $\mu_{\kappa(c)}$ over T epochs with $\log(K/\delta)$ replaced by $\log(K(T+1)/\delta)$.

Corollary 4.2 (Soft targets). *For a sample i with $c(i) = c$, let $\gamma_i = 1 - q_{i\kappa(c)}$ and $D_c = \max_{k \neq \kappa(c)} \|\mu_k - \nu_c\|_2$. Then $\|\bar{s}_i - \nu_c\|_2 \leq \|\mu_{\kappa(c)} - \nu_c\|_2 + \gamma_i D_c$, and for a poisoned sample and any unit vector u,*

$$|\langle s_i - \nu_c, u\rangle| \leq r_i\big(\|\mu_{\kappa(c)} - \nu_c\|_2 + \gamma_i D_c\big) + (1 - r_i)\big(|\langle \xi_i, u\rangle| + |\langle \delta_i, u\rangle|\big).$$

Proposition 4.3 (Alignment bounds the trigger-direction component). *For any unit vector $u \in \mathbb{R}^8$, if $1 - \cos(z_i^q, s_i) \leq \epsilon$, then $|\langle z_i^q, u\rangle| \leq \|z_i^q\|_2\big(|\langle s_i, u\rangle|/\|s_i\|_2 + \sqrt{2\epsilon}\big)$.*

Implications. Because Eqs. (1)–(2) never read labels, relabeling cannot move a prototype: a poisoned sample affects the mixture only through its latent offset δ_i, and both label mappings present identical inputs to the prototype branch. Proposition 4.1 bounds the deviation of the matched mean by a finite-sample term, a contamination term $\eta_\kappa\beta$, and a leakage term $\gamma\Delta$. Corollary 4.2 shows that a confidently assigned poisoned sample receives a target near the clean centroid, whereas an ambiguous one retains a fraction $1 - r_i$ of its own offset; empirically, the offset lowers the mean trust score from 0.791 (clean) to 0.694 (poisoned; Tab. 4a), cutting the cross-entropy weight of poisoned samples by 12% without removing any of them. Proposition 4.3 transfers the small trigger-direction mass of the target to the embedding that the classifier reads, so a classifier on z_i^q can realize the trigger rule only through a component controlled by the alignment residual. None of these results is a certified robustness guarantee; each identifies the quantity that controls the corresponding effect.

Why a tree-structured quantum encoder. The TQE gives the loop bounded outputs of the prototype's dimension, a seven-parameter graph whose Rademacher complexity cannot exceed that of any circuit class containing it (Proposition A.1), and a scale-invariant interface: $\mathcal{L}_{\mathrm{CL}}$ and SS_t are unchanged when equal-strength depolarizing or symmetric readout noise contracts every $\langle Z_j\rangle$ uniformly (Proposition A.3). A classical encoder with the same graph and budget falls 3.4 ACC points short (Tab. 3), so topology alone does not explain the contribution; we do not claim quantum advantage.

Failure mode. The bounds assume that poisoned samples share a component with their semantic class. A trigger offset large relative to class separation can instead form its own component, which violates (A3), is detected by Eq. (6) only through collapse or imbalance, and is the target of adaptive attacks on latent separability (Qi et al., 2023).

5 EXPERIMENTS

We evaluate five axes under one shared protocol: clean utility (Tab. 1 and App. Tab. 9), trigger suppression (Fig. 3), geometry (Figs. 4–5), mechanism attribution (Tabs. 2–4), and robustness beyond the default trigger (Tab. 5).

Table 1: **Clean accuracy (ACC, %, ↑) on OASIS and D1–D3** under random-class and targeted patch poisoning at $\eta \in \{1, 5, 10\}\%$. Mean±SD over three runs; best per column in bold. App. Tab. 9 reports D4–D6, App. D.2 macro-F1 and AUC, and App. D.3 the ASR of every method.

	OASIS			D1			D2			D3		
Method	1	5	10	1	5	10	1	5	10	1	5	10
Random-class patch attack												
AAMC	83.8±0.3	83.4±0.3	81.5±0.4	82.1±0.3	81.8±0.1	79.9±0.4	83.4±0.3	82.4±0.4	81.4±0.4	78.6±0.3	77.3±0.4	76.6±0.4
QcQN	85.7±0.3	85.3±0.3	83.3±0.4	84.0±0.3	83.6±0.3	81.7±0.4	85.2±0.3	84.2±0.3	83.2±0.4	80.4±0.3	79.0±0.4	78.3±0.4
Q-Detection	86.6±0.3	86.2±0.3	84.2±0.4	84.9±0.1	84.5±0.3	82.5±0.4	86.1±0.3	85.1±0.3	84.1±0.4	81.2±0.3	79.8±0.4	79.2±0.4
QuAdRo	84.8±0.3	84.4±0.3	82.4±0.4	83.1±0.3	82.7±0.4	80.8±0.4	84.3±0.3	83.3±0.4	82.3±0.4	79.5±0.3	78.1±0.7	77.5±0.4
QUID	88.1±0.3	87.9±0.3	86.1±0.4	85.7±0.3	86.0±0.3	83.3±0.4	86.9±0.3	86.1±0.3	85.1±0.4	82.9±0.3	80.5±0.4	80.1±0.4
NQNN	89.1±0.3	88.7±0.3	87.4±0.4	87.4±0.3	87.0±0.3	85.7±0.4	86.6±0.3	85.6±0.3	84.6±0.4	82.5±0.3	83.0±0.4	82.3±0.4
TSBD	87.5±0.3	87.1±0.3	85.1±0.4	85.8±0.3	85.4±0.3	83.4±0.4	87.0±0.3	86.0±0.3	85.0±0.4	82.1±0.2	80.7±0.4	80.0±0.4
DivideMix	89.3±0.3	88.9±0.3	87.1±0.3	84.5±0.3	87.1±0.3	86.2±0.4	88.8±0.3	87.7±0.3	87.1±0.4	83.9±0.3	82.2±0.4	81.5±0.4
Co-teaching	89.4±0.3	88.9±0.3	86.9±0.4	87.6±0.3	87.2±0.3	85.2±0.4	88.8±0.3	87.8±0.3	86.8±0.4	83.8±0.3	82.4±0.4	81.7±0.4
SAU	90.3±0.3	89.8±0.3	88.4±0.4	88.5±0.3	88.0±0.3	86.7±0.4	90.0±0.3	88.9±0.3	87.7±0.4	85.1±0.3	83.3±0.4	83.1±0.4
NP	91.2±0.3	90.8±0.3	88.6±0.4	89.4±0.3	89.0±0.3	86.9±0.4	90.7±0.3	89.6±0.3	88.6±0.4	85.5±0.3	84.1±0.3	83.4±0.4
TAPQ-Net	**92.1±0.3**	**91.7±0.3**	**89.6±0.4**	**90.3±0.3**	**89.9±0.3**	**87.8±0.4**	**91.6±0.3**	**90.5±0.3**	**89.5±0.4**	**86.4±0.3**	**84.9±0.3**	**84.2±0.4**
Targeted patch attack												
AAMC	82.2±0.4	81.8±0.4	79.9±0.5	80.6±0.4	80.2±0.4	78.3±0.5	81.7±0.4	80.8±0.4	79.8±0.5	77.0±0.4	75.8±0.4	75.1±0.5
QcQN	83.9±0.3	83.5±0.4	81.7±0.4	82.3±0.4	81.9±0.4	80.1±0.4	83.5±0.3	82.5±0.4	81.5±0.4	78.8±0.4	77.4±0.4	76.7±0.5
Q-Detection	84.9±0.3	84.5±0.4	82.5±0.4	83.2±0.4	82.8±0.4	80.9±0.4	84.4±0.3	83.4±0.4	82.4±0.4	79.6±0.4	78.2±0.4	77.6±0.5
QuAdRo	83.0±0.4	82.6±0.4	80.8±0.4	81.4±0.4	81.0±0.4	79.2±0.5	82.6±0.4	81.6±0.4	80.7±0.4	77.9±0.4	76.5±0.4	76.0±0.5
QUID	86.7±0.3	86.3±0.4	84.3±0.4	85.0±0.3	84.6±0.4	82.6±0.4	86.1±0.3	85.2±0.4	84.2±0.4	81.2±0.4	79.9±0.4	79.2±0.5
NQNN	87.4±0.3	87.0±0.4	85.7±0.4	85.7±0.3	85.3±0.4	84.0±0.4	84.9±0.3	83.9±0.4	82.9±0.4	80.9±0.4	81.3±0.4	80.7±0.4
TSBD	85.8±0.3	85.4±0.4	83.3±0.4	84.1±0.3	83.7±0.4	81.7±0.4	85.3±0.3	84.3±0.4	83.3±0.4	80.5±0.4	79.1±0.4	78.4±0.5
DivideMix	88.4±0.3	88.0±0.4	86.0±0.4	86.7±0.3	86.3±0.4	84.3±0.4	88.0±0.3	86.9±0.4	85.9±0.4	83.0±0.4	81.5±0.4	80.9±0.4
Co-teaching	87.5±0.3	87.2±0.4	85.2±0.4	85.8±0.3	85.5±0.4	83.5±0.4	87.0±0.3	86.0±0.4	85.1±0.4	82.1±0.4	80.8±0.4	80.1±0.4
SAU	88.4±0.3	87.9±0.4	86.7±0.4	86.7±0.3	86.2±0.4	85.0±0.4	88.2±0.3	87.1±0.4	86.0±0.4	83.4±0.3	81.6±0.4	81.4±0.4
NP	89.4±0.3	88.9±0.4	86.9±0.4	87.6±0.3	87.2±0.4	85.2±0.4	88.9±0.3	87.8±0.4	86.8±0.4	83.8±0.3	82.4±0.4	81.7±0.4
TAPQ-Net	**90.3±0.3**	**89.9±0.4**	**87.7±0.4**	**88.5±0.3**	**88.1±0.4**	**86.0±0.4**	**89.8±0.3**	**88.7±0.4**	**87.7±0.4**	**84.7±0.3**	**83.2±0.4**	**82.5±0.4**

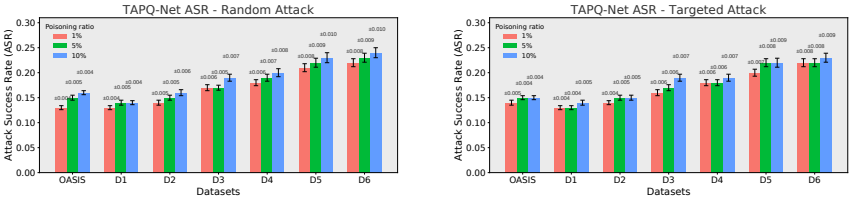
Datasets and attacks. We use OASIS MRI (Marcus et al., 2007) and six MedMNISTv2 datasets (Yang et al., 2023a): BloodMNIST (D1), PathMNIST (D2), OrganAMNIST (D3), OrganCMNIST (D4), OCTMNIST (D5), and DermaMNIST (D6), with 4–11 classes across six modalities (App. H.1). We split OASIS by patient before slice extraction and poisoning, so no patient spans two partitions. Following BadNets and BackdoorBench (Gu et al., 2017; Wu et al., 2022), we stamp a 3×3 white patch in the bottom-right corner of a fraction $\eta \in \{1, 5, 10\}\%$ of training images before normalization and relabel them by the random-class or targeted mapping, keeping validation and test images clean; on PathMNIST we also run the clean-label SIG and the warping-based WaNet attacks (Barni et al., 2019; Nguyen & Tran, 2021).

Metrics and statistics. We measure ACC, macro-F1, and AUC on clean test images. For targeted poisoning, ASR is the fraction of triggered test images with $y \neq t$ predicted as t; for random-class poisoning, each triggered test image x receives a seed-fixed incorrect label $t(x)$, and ASR is the fraction with $\widehat{y}(T(x)) = t(x)$. SS is the silhouette coefficient of the quantum embeddings (App. H.3). Values are means over three runs (full grids in App. D); for inference we pair the 42 conditions and apply Wilcoxon signed-rank tests with Holm correction (App. D.7).

Compared methods. We compare eleven methods grouped by design purpose: backdoor defenses (Neural Polarizer, NP (Zhu et al., 2023); SAU (Wei et al., 2023); TSBD (Lin et al., 2024)), noisy-label learning (DivideMix (Li et al., 2020); Co-teaching (Han et al., 2018)), and robustness references adapted to poisoned training (AAMC (Tsai et al., 2023); QcQN (Huang et al., 2022a); QuAdRo (Sahdev & Kumar, 2023); Q-Detection (He et al., 2025); QUID (Kundu & Ghosh, 2025); NQNN (Rahman & Zhuang, 2026a)). QUID is a data-poisoning attack and serves only as a poisoned-training reference. All methods share splits, poisoned indices, trigger, preprocessing, and evaluation (App. F); TAPQ-Net uses Adam (10^{-3}), batch size 128, and at most 15 epochs (App. H).

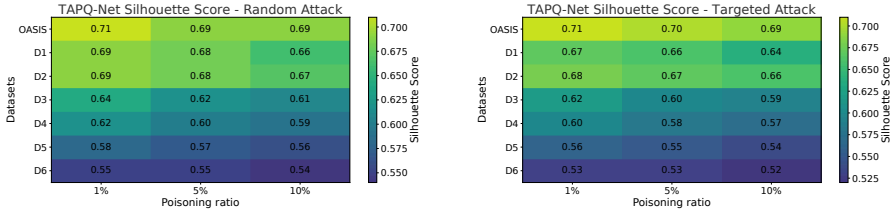
Trigger suppression. TAPQ-Net has the lowest ASR in all 42 conditions, from 0.13 to 0.24 (Fig. 3), whereas every compared method stays at or above 0.34 in every condition (App. Tab. 12). The lowest-ASR competitor is AAMC in every condition (0.34–0.46), and TAPQ-Net undercuts it by 0.21–0.22; the three backdoor defenses reach 0.36–0.53. The two label mappings change TAPQ-Net's ASR by at most 0.01, as expected from a mapping-blind prototype branch, and raising η from 1% to 10% increases it by at most 0.03 in every dataset–mapping pair.

Clean utility and significance. TAPQ-Net has the highest mean ACC in 40 of 42 conditions (Tab. 1 and App. Tab. 9); SAU leads by 0.2 points only on D5 at $\eta = 1\%$, and elsewhere the margin is 0.8–1.0 points against run-level standard deviations of 0.3–0.5. Macro-F1 is highest in all 42 conditions and AUC in 40, with NP ahead on OASIS under random-class poisoning at $\eta \in \{1, 5\}\%$



(a) Random-class patch attack. (b) Targeted patch attack.

Figure 3: **Attack success rate (ASR) of TAPQ-Net** on OASIS and D1–D6 at $\eta \in \{1, 5, 10\}\%$. Bars show means over three runs; error bars and the labels above them show standard deviations; lower is better. App. Tab. 12 compares all methods.



(a) Random-class patch attack. (b) Targeted patch attack.

Figure 4: **Silhouette coefficient of TAPQ-Net quantum embeddings.** Rows are OASIS and D1–D6; columns are $\eta \in \{1, 5, 10\}\%$; higher values indicate more compact and better-separated embeddings.

(App. Tabs. 10–11). Paired against NP, the strongest comparator overall, TAPQ-Net improves ACC by 0.85 points, macro-F1 by 0.82, and AUC by 0.82, and reduces ASR by 0.277; all four differences are significant after Holm correction (adjusted $p < 2 \times 10^{-7}$, App. Tab. 15). The advantage of TAPQ-Net is therefore concentrated in trigger suppression: its clean-accuracy margin over NP is at most 1.0 point, whereas its ASR is 0.21–0.22 below that of the lowest-ASR competitor in every condition.

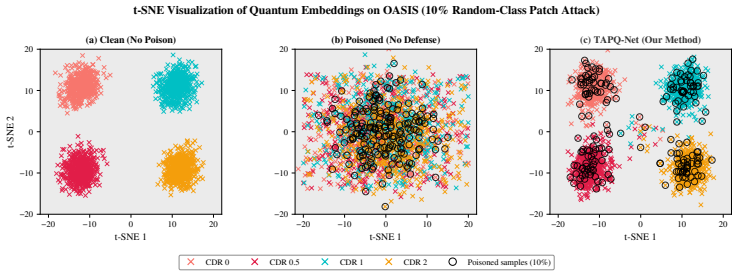


Figure 5: **t-SNE of quantum embeddings on OASIS** under random-class poisoning at $\eta = 10\%$ for clean training, undefended poisoned training, and TAPQ-Net. Colors denote CDR-based classes; black circles mark poisoned samples.

Embedding geometry. The silhouette coefficient of the quantum embedding remains positive in all 42 conditions (Fig. 4); on OASIS, TAPQ-Net forms more compact, class-consistent clusters than undefended poisoned training while retaining residual overlap between adjacent CDR stages (Fig. 5).
Geometry and residual attack success. The residual ASR of TAPQ-Net follows the geometry of its quantum embedding. Computed from the reported means, the silhouette coefficient (Fig. 4) and the ASR (Fig. 3) have a Spearman rank correlation of -0.92 across the 42 conditions (-0.95 under random-class and -0.93 under targeted poisoning): OASIS, D1, and D2, whose embeddings are best separated (mean SS 0.67–0.70), have ASR of at most 0.15, whereas D5 and D6, the least separated (0.54–0.56), have the highest ASR (0.22–0.23). This ordering is what Proposition 4.1 and Corollary 4.2 predict: weaker class separation enlarges the leakage term $\gamma\Delta$ and makes assignments less decisive, so soft targets retain a larger share $1 - r_i$ of the trigger offset. Because SS and ASR are measured on the same models and the datasets differ in difficulty, the correlation is consistent with the analysis rather than a test of it.

Component ablation. Tab. 2 reveals a staged mechanism rather than a sum of parts. The TQE alone (b) raises ACC over the linear classifier (a) but not trigger resistance (ASR 0.37 vs. 0.36). Prototype-weighted cross-entropy (c) is the only single component that lowers ASR (to 0.28); adding

Table 2: **Component ablation** on PathMNIST, random-class, $\eta = 10\%$. "Proto." enters only through the trust weights $\bar{r}_i$; "Align." adds $\mathcal{L}_{\mathrm{CL}}$; "Refresh" adds Eq. (6), without which the EM state stays fixed after initialization. SS is computed on the representation the classifier reads; Params follow Tab. 8.

	Variant	TQE	Proto.	Align.	Refresh	ACC↑	F1↑	AUC↑	ASR↓	SS↑	Params
(a)	Linear on z_i^c (no defense)	–	–	–	–	63.3	60.8	69.1	0.36	0.47	81
(b)	TQE + CE	✓	–	–	–	67.4	64.7	73.5	0.37	0.50	88
(c)	Proto-weighted CE on z_i^c	–	✓	–	–	72.5	69.5	79.0	0.28	0.53	171
(d)	TQE + proto-weighted CE	✓	✓	–	–	78.8	75.6	88.8	0.27	0.61	178
(e)	(d) + alignment	✓	✓	✓	–	80.6	77.3	87.8	0.29	0.60	178
(f)	**Full**: (e) + refresh	✓	✓	✓	✓	**89.5**	**85.9**	**97.6**	**0.16**	**0.67**	178

the TQE (d) raises ACC by 6.3 points and SS by 0.08. Aligning to a prototype state frozen after initialization (e) adds 1.8 ACC points but no ASR reduction, whereas feedback-controlled refresh (f) adds 8.9 ACC points, lowers ASR from 0.29 to 0.16, and raises SS from 0.60 to 0.67 at unchanged parameter count. Because the gate also controls how many EM passes occur, this row does not yet separate gating from continued refinement (App. D.6).

Table 3: **Encoder controls** on PathMNIST, random-class, $\eta = 10\%$; only the encoder changes. Params count encoder and classifier weights (TQE-8Q: $7 + 81$). The lower block is an independent rerun with a topology- and parameter-matched classical tree (App. D.8).

Encoder	ACC↑	F1↑	AUC↑	ASR↓	SS↑	Params
Encoder family and capacity (means over three runs)						
MLP-8D	74.9	71.6	82.4	0.235	0.54	88
Flat VQC-8Q	77.1	73.9	85.1	0.218	0.56	88
TQE-4Q	81.0	77.5	89.1	0.195	0.59	84
TQE-8Q, deeper	83.2	79.9	90.9	0.186	0.60	104
TQE-8Q	**89.5**	**85.9**	**97.6**	**0.160**	**0.67**	88
Topology-matched rerun (mean ± SD over three runs)						
Classical tree, same graph	85.21±0.24	–	–	0.191±0.004	0.633±0.005	88
TQE-8Q	**88.59±0.18**	–	–	**0.157±0.003**	**0.664±0.003**	88

Encoder controls. Inside the full loop, a parameter-matched MLP and a flat eight-qubit circuit fall 14.6 and 12.4 ACC points below TQE-8Q and raise ASR by 0.075 and 0.058; a four-qubit and a deeper eight-qubit tree are worse on every metric (Tab. 3). A classical tree with the same graph and 88-parameter budget trails the TQE by 3.38 ACC points, 0.034 ASR, and 0.031 SS (App. D.8).

Trust scores. Poisoned samples receive lower trust scores than clean ones (median 0.721 vs. 0.841), and r_i ranks clean above poisoned samples with AUROC 0.724 (Tab. 4a, App. D.9); poison annotations are consulted only after training. The overlap is expected, because r_i interpolates targets and weights and is never thresholded.

Compression. ACC, ASR, and SS improve monotonically with retained variance (Tab. 4b), and the step from $d_z = 6$ to 8 (+1.16 ACC) is a quarter of the step from 4 to 6 (App. D.10). Eight components retain 93.2% of the feature variance, so $d_z = 8$ is an operating point near saturation rather than a lossless projection. We restrict the sweep to $d_z \leq 8$ because larger representations would require additional qubits, grouped encoding, or data re-uploading, thereby changing the quantum architecture as well as the compression level.

Trigger variants. Tab. 5a varies the size, location, and opacity of the patch. TAPQ-Net degrades monotonically from the default trigger to the alpha-blended one, losing 4.3 ACC points while ASR rises from 0.160 to 0.228, which is consistent with larger or less coherent latent offsets enlarging the contamination term of Proposition 4.1. Over the same sequence, SS decreases from 0.67 to 0.58, showing that higher attack success coincides with weaker separation in the learned embedding.

Clean-label and warping attacks. Tab. 5b evaluates triggers that differ from the patch in kind rather than in size or location. Relative to NQNN, TAPQ-Net lowers ASR by 0.120 under SIG and by 0.129 under WaNet while raising ACC by 2.20 and 2.08 points (App. D.11). Because SIG leaves every label intact, this gain cannot come from down-weighting relabeled samples; it reflects the geometry of the latent offsets.

Table 4: **Trust and compression diagnostics** on PathMNIST, random-class, $\eta = 10\%$. (a) Poison annotations group the learned trust score r_i only after training. (b) Independent rerun, mean±SD over three runs; for $d_z < 8$, z_i^c is zero-padded so that the same eight-qubit TQE is used.

(a) Trust score r_i by group (AUROC 0.724).

Group	n	Mean±SD	Med.	IQR
Clean	80,996	0.791±0.141	0.841	0.754–0.908
Poisoned	9,000	0.694±0.187	0.721	0.593–0.835

(b) PCA dimension d_z.

d_z	Var. (%)	ACC↑	ASR↓	SS↑
2	57.3	77.15±0.53	0.245±0.012	0.543±0.009
4	75.6	82.48±0.41	0.206±0.009	0.594±0.007
6	87.6	86.99±0.31	0.173±0.006	0.641±0.005
8	**93.2**	**88.15±0.19**	**0.157±0.003**	**0.661±0.003**

Table 5: **Robustness beyond the default trigger** on PathMNIST at a 10% poisoning rate. (a) Patch variants under random-class poisoning. (b) Clean-label SIG and warping-based WaNet attacks, mean±SD over three runs.

(a) Patch size, location, and opacity.

Trigger	ACC↑	ASR↓	SS↑
3×3 corner (default)	**89.5**	**0.160**	**0.67**
3×3 center	88.9	0.183	0.66
5×5 corner	88.0	0.192	0.64
5×5 center	87.3	0.201	0.62
3×3, random location	86.1	0.214	0.60
3×3, alpha-blended	85.2	0.228	0.58

(b) Clean-label and warping attacks.

Attack	Method	ACC↑	ASR↓
SIG (clean-label)	NQNN	84.64±0.37	0.347±0.021
	TAPQ-Net	**86.84±0.28**	**0.227±0.015**
WaNet (warping)	NQNN	83.98±0.40	0.410±0.024
	TAPQ-Net	**86.06±0.32**	**0.281±0.018**

Backbones and cost. Four alternative frozen backbones change ACC by at most 1.4 points, ASR by at most 0.004, and SS by at most 0.011 per dataset (App. D.4). EM and gating run only during training, and inference takes 2.65 ms per image (App. D.5). Because the PCA dimension, EM branch, TQE, alignment objective, gate, and training protocol remain fixed in the backbone study, the narrow ASR and SS ranges indicate that trigger suppression and embedding geometry are largely insensitive to the frozen feature extractor, whereas the larger ACC variation reflects backbone-dependent clean utility. Because trainable capacity is not matched in the cost study, the 39.6-s epoch time should be interpreted descriptively; TAPQ-Net updates only 178 method-specific parameters, and its 2.65-ms inference latency lies between those of the classical and simulated-QNN baselines.

Limitations and evidence boundary. Five boundaries remain. (i) *Gating versus continued refinement*: always-refresh and fixed-interval controls are needed to attribute row (f) of Tab. 2 to the gate rather than to additional EM passes (App. G.2). (ii) *Adaptive attacks*: a trigger optimized to form a compact, well-separated component targets the failure mode of Sec. 4 (Qi et al., 2023); SIG and WaNet cover one dataset and one comparator (App. G.3). (iii) *Execution*: we simulate the circuit; Proposition A.3 covers only uniform readout contraction (App. G.4). (iv) *Scope*: the classical-tree, trust, and PCA studies use PathMNIST at $\eta = 10\%$, and paired tests over conditions that share datasets support consistency, not independence. (v) *Model selection*: epoch selection uses a clean validation split, which a fully poisoned pipeline may not provide (App. G.5). The results are benchmark evidence, not clinical validation; App. G specifies the open controls and App. G.1 the answered ones.

6 Conclusion

TAPQ-Net couples label-free EM prototypes with a seven-parameter tree-structured quantum encoder through an asymmetric loop: prototypes supply contrastive-alignment targets and trust weights, and an embedding-quality monitor on the quantum geometry decides whether the prototypes may evolve. Because the prototypes never read labels, the attacker reaches them only through the trigger's latent offset (Proposition 4.1), and alignment bounds the trigger-direction component of the embedding (Proposition 4.3). Across 42 poisoning conditions on seven medical datasets, TAPQ-Net attains the lowest ASR everywhere and the highest clean accuracy in 40, outperforms a classical tree of identical topology and budget, and extends to clean-label and warping triggers, while its residual ASR tracks the separation of its embedding as the analysis predicts. The evidence also delimits the claim: the gains come from the loop rather than from a quantum advantage, and gate-ablation controls, adaptive triggers that form their own latent component, and execution under device noise are the next tests of the mechanism.

AI USE STATEMENT

We used OpenAI's ChatGPT to assist with language editing, manuscript organization, and LaTeX formatting. The authors verified all scientific claims, analyses, results, citations, and final text.

REPRODUCIBILITY STATEMENT

Sec. 3 and App. C define every component, including the refresh rule (Eq. (6), App. B) and the training procedure (Algorithms 1–2). App. H lists datasets, splits, the trigger operators, the ASR definitions, hyperparameters (Tab. 6), software versions, and hardware. Sec. 4 states all assumptions, App. A gives all proofs, and App. D.7 specifies the statistical tests.

REFERENCES

N. Ajlouni, A. Ozyavas, M. Takaoglu, F. Takaoglu, and F. Ajlouni. Medical image diagnosis based on adaptive hybrid quantum cnn. *BMC Medical Imaging*, 23(1):126, 2023.

Venkat Adithya Amula, Sunayana Samavedam, Saurabh Saini, Avani Gupta, and P. J. Narayanan. Prototype guided backdoor defense via activation space manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2195–2205, 2025.

Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in CNNs by training set corruption without label poisoning. In *IEEE International Conference on Image Processing (ICIP)*, pp. 101–105, 2019.

Marcello Benedetti, Erika Lloyd, Stefan Sack, and Mattia Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, 2019. doi: 10.1088/2058-9565/ab4eb5.

D. Bokhan, A. S. Mastiukova, A. S. Boev, D. N. Trubnikov, and A. K. Fedorov. Multiclass classification using quantum convolutional neural networks with hybrid quantum-classical learning. *Frontiers in Physics*, 10:1069985, 2022.

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, 2021. doi: 10.1038/s41467-021-21728-w.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. In *Workshop on Artificial Intelligence Safety (SafeAI)*, 2019.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Iris Cong, Soonwon Choi, and Mikhail D. Lukin. Quantum convolutional neural networks. *Nature Physics*, 15:1273–1278, 2019. doi: 10.1038/s41567-019-0648-8.

Bappaditya Dhara, Manan Agrawal, and Soumya D. Roy. Multi-class classification using quantum transfer learning. *Quantum Information Processing*, 23(2):34, 2024.

Junhao Dong, Junxi Chen, Xiaohua Xie, Jianhuang Lai, and Hao Chen. Survey on adversarial attack and defense for medical image analysis: Methods and challenges, 2024.

Yuxuan Du, Min-Hsiu Hsieh, Tongliang Liu, Dacheng Tao, and Nana Liu. Quantum noise protects quantum classifiers against adversaries. *Physical Review Research*, 3(2):023153, 2021.

Samuel G. Finlayson, John D. Bowers, Joichi Ito, Jonathan L. Zittrain, Andrew L. Beam, and Isaac S. Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019. doi: 10.1126/science.aaw4399.

Long Gao, Lei Zhang, Chang Liu, and Shandong Wu. Handling imbalanced medical image data: A deep-learning-based one-class classification approach. *Artificial Intelligence in Medicine*, 108: 101935, 2020. doi: 10.1016/j.artmed.2020.101935.

Edward Grant, Marcello Benedetti, Shuxiang Cao, Andrew Hallam, Joshua Lockhart, Vid Stojevic, Andrew G. Green, and Simone Severini. Hierarchical quantum classifiers. *npj Quantum Information*, 4(1):65, 2018. doi: 10.1038/s41534-018-0116-9.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

J. Guo et al. Backdoor attacks against hybrid classical-quantum neural networks. *Neural Networks*, 191:107776, 2025a. doi: 10.1016/j.neunet.2025.107776.

Wei Guo, Maura Pintor, Ambra Demontis, and Battista Biggio. Prototype-guided robust learning against backdoor attacks. *arXiv preprint arXiv:2509.08748*, 2025b.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Haoqi He, Xiaokai Lin, Jiancai Chen, and Yan Xiao. Q-detection: A quantum-classical hybrid poisoning attack detection method. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5327–5335, 2025. doi: 10.24963/ijcai.2025/593.

Jhih-Cing Huang, Yu-Lin Tsai, Chao-Han Huck Yang, Cheng-Fang Su, Chia-Mu Yu, Pin-Yu Chen, and Sy-Yen Kuo. Certified robustness of quantum classifiers against adversarial examples through quantum noise, 2022a.

Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In *International Conference on Learning Representations*, 2022b.

William Huggins, Piyush Patel, K. Birgitta Whaley, and E. Miles Stoudenmire. Towards quantum machine learning with tensor networks. *Quantum Science and Technology*, 4(2):024001, 2019. doi: 10.1088/2058-9565/aaea94.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

Rishi D. Jha, Jonathan Hayase, and Sewoong Oh. Label poisoning is all you need. In *Advances in Neural Information Processing Systems*, 2023.

Peng Jin, Haochen Liu, Trung Bui, Yu Chen, Zhaowen Zhou, Xiaojun Han, and Andrei Barbu. Expectation-maximization contrastive learning for compact video-and-language representations. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30291–30306, 2022.

Luyang Ju, Xin Wang, Le Wang, Dwarikanath Mahapatra, Xin Zhao, Quan Zhou, Tieniu Liu, and Zongyuan Ge. Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE Transactions on Medical Imaging*, 41(6):1533–1546, 2022.

Davood Karimi, Haili Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

Satwik Kundu and Swaroop Ghosh. Adversarial data poisoning attack on quantum machine learning in the nisq era. In *Proceedings of the Great Lakes Symposium on VLSI (GLSVLSI)*, pp. 976–981, 2025. doi: 10.1145/3716368.3735244.

Mohammed Kutbi. Effective defense against physically embedded backdoor attacks via clustering-based filtering. *Complex & Intelligent Systems*, 11:241, 2025. doi: 10.1007/s40747-025-01876-y.

Jonas Landman, Nishant Mathur, Yun Yao Li, Markus Strahm, Sofiene Kazdaghli, Aniket Prakash, and Iordanis Kerenidis. Quantum methods for neural networks and application to medical image classification. *Quantum*, 6:881, 2022.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021a.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. In *Advances in Neural Information Processing Systems*, pp. 14900–14912, 2021b.

Weilin Lin, Li Liu, Shaokui Wei, Jianze Li, and Hui Xiong. Unveiling and mitigating backdoor vulnerabilities based on unlearning weight changes and backdoor activeness. In *Advances in Neural Information Processing Systems*, 2024.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017. doi: 10.1016/j.media.2017.07.005.

Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (oasis): Cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9): 1498–1507, 2007. doi: 10.1162/jocn.2007.19.9.1498.

Nishant Mathur, Jonas Landman, Yun Yao Li, Markus Strahm, Sofiene Kazdaghli, Aniket Prakash, and Iordanis Kerenidis. Medical image classification via quantum neural networks, 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, 2018. doi: 10.1038/s41467-018-07090-4.

Rui Min, Zeyu Qin, Nevin L. Zhang, Li Shen, and Minhao Cheng. Uncovering, explaining, and mitigating the superficial safety of backdoor defense. In *Advances in Neural Information Processing Systems*, 2024.

Tuan Anh Nguyen and Anh Tuan Tran. WaNet – imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021.

Román Orús. A practical introduction to tensor networks: Matrix product states and projected entangled pair states. *Annals of Physics*, 349:117–158, 2014. doi: 10.1016/j.aop.2014.06.013.

Arthur Pesah, M. Cerezo, Samson Wang, Tyler Volkoff, Andrew T. Sornborger, and Patrick J. Coles. Absence of barren plateaus in quantum convolutional neural networks. *Physical Review X*, 11(4): 041011, 2021. doi: 10.1103/PhysRevX.11.041011.

Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, and Prateek Mittal. Revisiting the assumption of latent separability for backdoor defenses. In *International Conference on Learning Representations*, 2023.

Maqsudur Rahman and Jun Zhuang. Nqnn: Noise-aware quantum neural networks for medical image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15972 of *Lecture Notes in Computer Science*, pp. 433–442. Springer, 2026a. doi: 10.1007/978-3-032-05169-1_42.

Maqsudur Rahman and Jun Zhuang. QAPNet: A quantum-attentive patchwise network for robust medical image classification under noisy inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026b.

Sivaramakrishnan Rajaraman, Prasanth Ganesan, and Sameer Antani. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. *PLOS ONE*, 17 (1):e0262838, 2022. doi: 10.1371/journal.pone.0262838.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine*, 3(1):119, 2020. doi: 10.1038/s41746-020-00323-1.

Aditya Sahdev and Mehul Kumar. Adversarial robustness based on randomized smoothing in quantum machine learning. In *International Conference on Learning Representations Workshop/Submission*, 2023. URL `https://openreview.net/forum?id=o-Yxq5iicIp`.

Yucheng Shi, Mengnan Du, Xuansheng Wu, Zihan Guan, Jin Sun, and Ninghao Liu. Black-box backdoor defense via zero-shot image purification. In *Advances in Neural Information Processing Systems*, 2023.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Edwin Miles Stoudenmire and David J. Schwab. Supervised learning with quantum-inspired tensor networks. In *Advances in Neural Information Processing Systems*, 2016.

Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019. doi: 10.1038/s41591-018-0300-7.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, 2018.

Min-Jen Tsai, Ping-Yi Lin, and Ming-En Lee. Adversarial attacks on medical image classification. *Cancers*, 15(17):4228, 2023. doi: 10.3390/cancers15174228.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

S. Wang, Z. Xiao, J. Shi, H. Shi, S. Zhang, and X. Li. QSentry: Backdoor detection for quantum neural networks via measurement clustering. *arXiv preprint arXiv:2511.15376*, 2025.

Shaokui Wei, Mingda Zhang, Hongyuan Zha, and Baoyuan Wu. Shared adversarial unlearning: Backdoor mitigation by unlearning shared adversarial examples. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jiancheng Yang, Rui Shi, Donglai Wei, Zehan Liu, Lin Zhao, Bin Ke, Hanspeter Pfister, and Bing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023a.

Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Robust contrastive language-image pre-training against data poisoning and backdoor attacks. In *Advances in Neural Information Processing Systems*, 2023b.

Yanxin Yang, Chentao Jia, DengKe Yan, Ming Hu, Tianlin Li, Xiaofei Xie, Xian Wei, and Mingsong Chen. Sampdetox: Black-box backdoor defense via perturbation-based sample detoxification. In *Advances in Neural Information Processing Systems*, 2024.

M. Yousif, B. Al-Khateeb, and B. Garcia-Zapirain. A new quantum circuits of quantum convolutional neural network for x-ray images classification. *IEEE Access*, 2024.

Yulu Zhang. Research on detection and defense methods of backdoor attacks on quantum neural networks. *Scientific Reports*, 16:25961, 2026. doi: 10.1038/s41598-026-57090-4.

Zaixi Zhang, Qi Liu, Zhicai Wang, Zepu Lu, and Qingyong Hu. Backdoor defense via deconfounded representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12228–12238, 2023.

Mingli Zhu, Shaokui Wei, Hongyuan Zha, and Baoyuan Wu. Neural polarizer: A lightweight and effective backdoor defense via purifying poisoned features. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

APPENDIX OVERVIEW

This appendix supports the main paper in nine parts. App. A proves the results of Sec. 4 and adds two properties of the TQE; App. B gives the exact refresh rule and the hyperparameters; App. C collects the notation, the training algorithms, and the parameter accounting; App. D reports the complete evaluation grid, the paired significance analysis, and the diagnostics summarized in Sec. 5; App. E compares TAPQ-Net with the closest prior work at the level of mechanism; App. F describes the baseline protocol; App. G separates the answered controls from the open ones; App. H lists the datasets, configuration, attack protocol, and software; and App. I extends the related-work discussion.

A THEORETICAL ANALYSIS

This section proves the results of Sec. 4, whose setting, notation, and assumptions (A1)–(A4) we use throughout, and states two further properties of the TQE: Proposition A.1 relates its capacity to that of circuits containing it, and Proposition A.3 shows that the alignment loss and the refresh gate are invariant to uniform readout contraction.

A.1 PROOFS FOR PROTOTYPE DEVIATION AND SOFT TARGETS

Proof of Proposition 4.1. Because $\sum_i w_i = 1$, $\widehat{\mu}_\kappa - \nu_c = \sum_i w_i(z_i^c - \nu_c)$. Substituting (A1),

$$\widehat{\mu}_\kappa - \nu_c = \underbrace{\sum_{i \in \mathcal{C}_c \cup \mathcal{P}_c} w_i \xi_i}_{X} + \sum_{i \in \mathcal{P}_c} w_i \delta_i + \sum_{i \in \mathcal{O}_c} w_i (z_i^c - \nu_c).$$

By (A2) and $\|\delta_i\|_2 \le \beta$, the second term has norm at most $\eta_\kappa \beta$; by (A3), the third has norm at most $\gamma\Delta$. By (A4) the weights are fixed, so every projection $\langle X, u\rangle$ is a weighted sum of independent sub-Gaussian variables with variance proxy at most $\varsigma^2 \sum_i w_i^2 \le \varsigma^2 c_0 / n_c$. A standard $\frac{1}{2}$-net argument over the unit sphere of $\mathbb{R}^{d_z}$ gives $\|X\|_2 \le C_1 \varsigma \sqrt{c_0(d_z + \log(1/\delta'))/n_c}$ with probability at least $1 - \delta'$. Setting $\delta' = \delta/K$ and taking a union bound over the K matched components yields the bound for $\widehat{\mu}_{\kappa(c)}$. Each accepted EM pass replaces μ_κ by a convex combination of its previous value and $\widehat{\mu}_\kappa$ (Sec. 3); if the initial means satisfy the same bound, a union bound over at most $T + 1$ passes, with $\log(K/\delta)$ replaced by $\log(K(T+1)/\delta)$, extends the bound to the EMA-stabilized means by convexity of the norm. □

Proof of Corollary 4.2. Since $\sum_k q_{ik} = 1$, $\bar{s}_i - \nu_c = \sum_k q_{ik}(\mu_k - \nu_c)$, whose norm is at most $q_{i\kappa}\|\mu_\kappa - \nu_c\|_2 + \sum_{k \ne \kappa} q_{ik}\|\mu_k - \nu_c\|_2 \le \|\mu_\kappa - \nu_c\|_2 + \gamma_i D_c$. Eq. (3) and (A1) give the identity $s_i - \nu_c = r_i(\bar{s}_i - \nu_c) + (1 - r_i)\big(\xi_i + \mathbf{1}[i \in \mathcal{P}]\,\delta_i\big)$, and the projection bound follows from the triangle inequality. □

The corollary makes the trade-off of the soft target explicit: a poisoned sample confidently assigned to its matched component receives a target close to the clean centroid, and the trigger reaches the target only through $\eta_\kappa \beta$ in the mean and through the residual weight $1 - r_i$ on the sample's own offset. When triggered samples form their own component, (A3) fails for the affected classes; this is the failure mode of Sec. 4.

A.2 PROOF OF TRIGGER-DIRECTION ATTENUATION

Proof of Proposition 4.3. Let $\hat{z} = z_i^q / \|z_i^q\|_2$ and $\hat{s} = s_i / \|s_i\|_2$. Then $\|\hat{z} - \hat{s}\|_2^2 = 2 - 2\langle \hat{z}, \hat{s}\rangle \le 2\epsilon$. By the triangle and Cauchy–Schwarz inequalities, $|\langle \hat{z}, u\rangle| \le |\langle \hat{s}, u\rangle| + \|\hat{z} - \hat{s}\|_2 \le |\langle s_i, u\rangle| / \|s_i\|_2 + \sqrt{2\epsilon}$. Multiplying by $\|z_i^q\|_2$ gives the claim. □

Combined with Corollary 4.2, a small alignment residual transfers the limited trigger-direction mass of the target to the embedding read by the classifier. The residual that can be reached is architecture-dependent: by Remark A.2, four coordinates of z_i^q are fixed functions of z_i^c, so ϵ cannot be driven to zero unless s_i agrees with them in direction.

A.3 CAPACITY OF THE TREE-STRUCTURED ENCODER

Let $\mathcal{H}_{\mathrm{TQE}} = \{z \mapsto \langle w, z_\theta^q(z)\rangle + b : \theta \in \Theta,\ \|w\|_2 \le B_w,\ |b| \le B_b\}$, and let $\mathcal{H}_{\mathrm{VQC}}$ be the analogous class for a circuit whose gate sequence contains the TQE blocks together with additional parameterized blocks, each of which equals the identity for some admissible parameter value (for example, controlled rotations $CR_Y(\zeta)$ at $\zeta = 0$), with the same readout constraints. For a sample $S = \{z_i^c\}_{i=1}^{n_S}$ and independent Rademacher signs σ_i, $\widehat{\mathfrak{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma\big[\sup_{h\in\mathcal{H}} \frac{1}{n_S}\sum_i \sigma_i h(z_i^c)\big]$.

Proposition A.1 (Capacity inclusion)**.** *Under the identity-attainability assumption, $\mathcal{H}_{\mathrm{TQE}} \subseteq \mathcal{H}_{\mathrm{VQC}}$ and hence $\widehat{\mathfrak{R}}_S(\mathcal{H}_{\mathrm{TQE}}) \le \widehat{\mathfrak{R}}_S(\mathcal{H}_{\mathrm{VQC}})$ for every S. The inequality is strict if and only if some sign vector $\sigma \in \{\pm 1\}^{n_S}$ gives a strictly larger supremum over $\mathcal{H}_{\mathrm{VQC}}$.*

Proof. Setting every additional block to the identity reproduces any TQE function, so the inclusion holds. For each fixed σ, a supremum over a subset cannot exceed the supremum over the superset; averaging over σ gives the inequality. Since each of the 2^{n_S} sign vectors has probability 2^{-n_S}, the average is strictly larger if and only if at least one term is. □

The flat VQC-8Q of Tab. 3 is parameter-matched to the TQE and uses different connectivity; it is not a superset of the TQE, so Proposition A.1 does not apply to that comparison, which is empirical. Lower capacity also does not by itself imply robustness.

Remark A.2 (Leaf and path coordinates). For $j \in \{0, 2, 4, 6\}$, qubit Q_j appears only as the control of $\mathrm{CNOT}_{j\to j+1}$ in $\mathcal{P}_1$. Since $\mathrm{CNOT} = |0\rangle\langle 0| \otimes I + |1\rangle\langle 1| \otimes X$ commutes with Z on its control, and every other gate acts on other qubits, the Heisenberg-evolved observable is $U^\dagger Z_j U = Z_j$. Hence $\langle Z_j\rangle_i = \cos z_{i,j}^c$, a parameter-free feature. For Q_1, which is a target in $\mathcal{P}_1$ and only a control afterwards, $\mathrm{CNOT}_{0\to1}^\dagger Z_1 \mathrm{CNOT}_{0\to1} = Z_0 Z_1$ and X_1 commutes with the CNOT; the subsequent $R_Y(\theta_1^{(1)})$ rotates the Bloch vector of Q_1 about Y, giving

$$\langle Z_1\rangle_i = \cos\theta_1^{(1)} \cos z_{i,0}^c \cos z_{i,1}^c - \sin\theta_1^{(1)} \sin z_{i,1}^c,$$

and analogously for Q_5. The coordinates $\langle Z_3\rangle_i$ and $\langle Z_7\rangle_i$ depend on two and three trainable angles, respectively, and aggregate progressively larger subsets of the input coordinates. This is the sense in which the TQE combines a fixed cosine feature map with a learned local-to-global path.

A.4 INVARIANCE TO UNIFORM READOUT CONTRACTION

Proposition A.3 (Uniform contraction)**.** *Suppose that, immediately before measurement, every qubit undergoes either a single-qubit depolarizing channel $\Lambda_j(\rho) = (1-p)\rho + p\,\mathrm{Tr}_j(\rho)\otimes\frac{I}{2}$ with common $p \in [0, 1)$, or a symmetric readout bit flip with common probability $p \in [0, \frac{1}{2})$. Then the observed embedding is $\tilde{z}_i^q = \omega\, z_i^q$ with $\omega = 1 - p$ or $\omega = 1 - 2p$, respectively, and (i) $\mathcal{L}_{\mathrm{CL}}$ is unchanged, and so is its gradient with respect to θ; (ii) SS_t, and hence every decision of the gate in Eq. (6), is unchanged; (iii) the normalized bound of Proposition 4.3 is unchanged; (iv) the classifier logits become $\omega\, W_{\mathrm{cls}} z_i^q + b$, so predictions are unchanged when $b = 0$ and otherwise depend on ω.*

Proof. For depolarizing noise, $\mathrm{Tr}(Z_j\Lambda_j(\rho)) = (1-p)\mathrm{Tr}(Z_j\rho)$ because $\mathrm{Tr}(Z) = 0$; for $l \neq j$, Λ_l is trace-preserving, so its adjoint is unital and $\Lambda_l^\dagger(Z_j) = Z_j$. For readout flips, the expected outcome is $(1-p)\langle Z_j\rangle - p\langle Z_j\rangle = (1-2p)\langle Z_j\rangle$. Thus $\tilde{z}_i^q = \omega z_i^q$ with $\omega > 0$. Cosine similarity is invariant to positive rescaling of either argument, which gives (i) and (iii). Every pairwise distance scales by ω, so the mean within-component distance $d_{\mathrm{in}}(i)$ and the mean nearest-component distance $d_{\mathrm{out}}(i)$ both scale by ω, and each silhouette value $(d_{\mathrm{out}}(i) - d_{\mathrm{in}}(i))/\max\{d_{\mathrm{in}}(i), d_{\mathrm{out}}(i)\}$ is unchanged, which gives (ii). (iv) is immediate. □

The proposition does not cover noise between gates, which generally does not reduce to a uniform contraction; amplitude damping, which adds an offset; or finite-shot estimation, which adds variance of order $S^{-1/2}$ for S shots. These effects are not evaluated (Sec. 5).

B Refresh Rule and Hyperparameters

After every epoch t, the TQE is applied to the full poisoned training set and the following label-free quantities are computed from $\{z_i^q\}_{i=1}^N$ and the monitoring assignments $\{a_i\}_{i=1}^N$ of the current accepted state:

$$SS_t = \mathrm{Silhouette}\big(\{z_i^q\}, \{a_i\}\big), \qquad m_k^{(t)} = \sum_{i=1}^{N} \mathbf{1}[a_i = k],$$

$$\rho_t = \frac{\max_k m_k^{(t)}}{N}, \qquad K_{\mathrm{act}} = \sum_{k=1}^{K} \mathbf{1}\big[m_k^{(t)} \geq m_{\min}\big].$$

With $SS_{\mathrm{best}}^{(t-1)} = \max_{1 \leq u < t} SS_u$ (initialized to $-\infty$), a refresh is executed only if

$$t \bmod \Delta_{\mathrm{ref}} = 0, \qquad SS_t \geq SS_{\mathrm{best}}^{(t-1)} - \varepsilon_{\mathrm{sil}}, \qquad \rho_t \leq \rho_{\max}, \qquad K_{\mathrm{act}} \geq K_{\min}. \tag{7}$$

This is Eq. (6) of the main paper. The silhouette criterion rejects a refresh when the quantum geometry has degraded relative to its best value; $\rho_{\max}$ rejects a state in which one component dominates; and $K_{\min}$ rejects a state in which too few components remain populated. If any condition fails, the previous state, including $\{s_i\}$, $\{r_i\}$, and $\{a_i\}$, is retained. The threshold $m_{\min}$ is applied to the soft mass M_k inside the EM update (Algorithm 1) and to the hard count $m_k^{(t)}$ in the gate. Tab. 6 lists all hyperparameters.

Table 6: **Hyperparameters of TAPQ-Net** and the equations in which they act. One setting is shared across datasets, label mappings, and poisoning ratios; Tab. 19 lists the training configuration.

Symbol	Acts in	Role
$d_z = 8$	Sec. 3	PCA latent dimension, equal to the number of qubits
$K = C$	Eq. (1)	number of mixture components
τ	Eq. (1)	E-step temperature; larger values soften the responsibilities
α	Sec. 3	EMA factor of each accepted EM pass
$m_{\min}$	Sec. 3, Eq. (6)	minimum component support for an EMA update and for an active component
λ_{CL}	Eq. (5)	weight of the prototype–quantum contrastive alignment loss
$r_{\min}$	Eq. (5)	floor of the trust weight, which keeps every sample in the objective
Δ_{ref}	Eq. (6)	refresh interval in epochs
$\varepsilon_{\mathrm{sil}}$	Eq. (6)	tolerated silhouette decrease relative to the best value so far
$\rho_{\max}$	Eq. (6)	maximum share of samples assigned to one component
$K_{\min}$	Eq. (6)	minimum number of active components

C Method Details

C.1 Notation

Tab. 7 summarizes the notation. Semantic classes are indexed by (c, C) and prototype components by (k, K). The superscripts in z_i^c and z_i^q denote the classical and quantum branches and are not class indices.

C.2 Algorithms

Algorithm 1 is one EM pass over the full latent set; it is run once at initialization and once at every accepted refresh. Algorithm 2 combines it with mini-batch optimization of the TQE and classifier and with the gate of Eq. (7).

C.3 Parameter accounting

For C classes, the gradient-trained parameters are the seven TQE angles and the $8C+C$ weights of the linear classifier; the EM state contributes $K(d_z + 2)$ estimated parameters (d_z mean coordinates, one variance, and one weight per component). With $C = K = 9$ (PathMNIST), this gives $7 + 81 + 90 =$

Table 7: **Principal notation.**

Symbol	Meaning
$x_i,\ y_i,\ \mathcal{Y}$	input image, observed (possibly poisoned) label, label set $\{1,\dots,C\}$
$\mathcal{P},\ \eta,\ T,\ t$	poisoned index set, poisoning ratio, trigger operator, all-to-one target
$f_i,\ W,\ \bar{f}$	frozen feature vector, PCA projection, feature mean of the training split
$z_i^c,\ d_z$	PCA latent vector, $d_z = 8$
z_i^q	TQE embedding $[\langle Z_0\rangle_i,\dots,\langle Z_7\rangle_i]^\top \in [-1,1]^8$
$\theta_w^{(\ell)},\ \mathcal{P}_\ell$	trainable rotation on target Q_w at level ℓ; interaction set of level ℓ
$W_{\mathrm{cls}},\ b$	linear classifier on z_i^q
$\pi_k,\ \mu_k,\ v_k,\ M_k$	mixture weight, mean, isotropic variance, and responsibility mass of component k
$\tau,\ \alpha,\ m_{\min}$	E-step temperature, EMA factor, minimum component support
$q_{ik},\ r_i,\ a_i$	responsibility, trust score $\max_k q_{ik}$, monitoring assignment $\arg\max_k q_{ik}$
$\bar{s}_i,\ s_i$	mixture target $\sum_k q_{ik}\mu_k$ and soft target $r_i\bar{s}_i + (1-r_i)z_i^c$
$\mathcal{L}_{\mathrm{CE}},\ \mathcal{L}_{\mathrm{CL}},\ \lambda_{\mathrm{CL}},\ \bar{r}_i,\ r_{\min}$	weighted cross-entropy, alignment loss, alignment weight, clipped trust score, floor
$SS_t,\ \rho_t,\ K_{\mathrm{act}},\ SS_{\mathrm{best}}$	epoch silhouette, largest-component ratio, active components, best silhouette so far
$\Delta_{\mathrm{ref}},\ \varepsilon_{\mathrm{sil}},\ \rho_{\max},\ K_{\min}$	refresh interval and gate thresholds (Eq. (6))
$\nu_c,\ \varsigma,\ \delta_i,\ \beta$	clean centroid, sub-Gaussian parameter, trigger latent offset and its bound (Sec. 4)
$\kappa(c),\ \eta_\kappa,\ \gamma,\ \Delta,\ n_c$	matched component, contamination mass, leakage mass, leakage radius, class size (Sec. 4)

Algorithm 1 One EM pass (initialization and every accepted refresh).

Input: latent vectors $\{z_i^c\}_{i=1}^N$; state $\{\pi_k,\mu_k,v_k\}_{k=1}^K$; τ, α, $m_{\min}$.
Output: updated state; $\{q_{ik}\}$, $\{r_i\}$, $\{a_i\}$, $\{s_i\}$.
for $i = 1,\dots,N$ **do**
  Compute ℓ_{ik} for all k and q_{ik} by Eq. (1).
  $r_i \leftarrow \max_k q_{ik}$; $a_i \leftarrow \arg\max_k q_{ik}$.
end for
for $k = 1,\dots,K$ **do**
  $M_k \leftarrow \sum_i q_{ik}$; compute $\widehat{\pi}_k, \widehat{\mu}_k, \widehat{v}_k$ by Eq. (2).
  $\pi_k \leftarrow (1-\alpha)\pi_k + \alpha\widehat{\pi}_k$.
  if $M_k \geq m_{\min}$ **then** $\mu_k \leftarrow (1-\alpha)\mu_k + \alpha\widehat{\mu}_k$; $v_k \leftarrow (1-\alpha)v_k + \alpha\widehat{v}_k$.
end for
for $i = 1,\dots,N$ **do** $\bar{s}_i \leftarrow \sum_k q_{ik}\mu_k$; $s_i \leftarrow r_i\bar{s}_i + (1-r_i)z_i^c$ **end for**

Algorithm 2 TAPQ-Net training with feedback-controlled prototype refresh.

Input: poisoned training set $\mathcal{D}^p$; frozen encoder; epochs T; hyperparameters of Tab. 6.
Extract $\{f_i\}_{i=1}^N$; fit PCA on $\{f_i\}$ once; compute $\{z_i^c\}_{i=1}^N$ and fix the projection.
Initialize $\{\mu_k\}$ by clustering $\{z_i^c\}$ into K groups; initialize $\{\pi_k\}$ and $\{v_k\}$.
Run Algorithm 1 to obtain the initial state $\mathcal{S}_0$; set $SS_{\mathrm{best}} \leftarrow -\infty$.
for $t = 1,\dots,T$ **do**
  for each mini-batch $\mathcal{B}$ **do**
    Read z_i^c, s_i, r_i from $\mathcal{S}_{t-1}$; compute z_i^q by the TQE and $\widehat{y}_i$ by Eq. (4).
    Update θ and (W_{cls}, b) by a gradient step on $\mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{CL}}\mathcal{L}_{\mathrm{CL}}$ (Eq. (5)).
  end for
  Compute $\{z_i^q\}_{i=1}^N$ with the updated TQE; compute SS_t, ρ_t, K_{act} using $\{a_i\}$ of $\mathcal{S}_{t-1}$.
  if Eq. (7) holds **then** $\mathcal{S}_t \leftarrow$ Algorithm 1 applied to $\mathcal{S}_{t-1}$ **else** $\mathcal{S}_t \leftarrow \mathcal{S}_{t-1}$.
  $SS_{\mathrm{best}} \leftarrow \max(SS_{\mathrm{best}}, SS_t)$.
end for
Return the TQE and classifier of the epoch selected on the validation split.

178 (Tab. 8). The same decomposition produces the counts of Tab. 2: 81 for the linear classifier (row a), $7 + 81 = 88$ with the TQE (row b), $81 + 90 = 171$ with the prototype state and no TQE (row c), and 178 for rows d–f. In Tab. 3, the counts exclude the 90 EM parameters shared by all rows. Frozen backbones and the PCA projection are excluded throughout.

Table 8: **Method-specific parameters** for $C = 9$ (PathMNIST). The frozen encoder and the fixed PCA projection are excluded.

Component	Fitted by	Count
TQE rotations θ	gradient	$4{+}2{+}1 = 7$
Classifier (W_{cls}, b)	gradient	$8C{+}C = 81$
Prototypes $\{\pi_k, \mu_k, v_k\}$	EM	$K(d_z{+}2) = 90$
Total		**178**

D Additional Results

This section completes the evaluation grid of Sec. 5. Tab. 1 and Fig. 3 in the main text report OASIS and D1–D3 accuracy and the ASR of TAPQ-Net with standard deviations; here we add the D4–D6 accuracy, macro-F1, AUC, and ASR for every method, the paired significance analysis, the diagnostics behind Tabs. 3–5, the backbone study, and the computational cost. All values are means over three independent runs; the dense F1, AUC, and ASR tables omit standard deviations for legibility. The best value in each column is in bold. Baselines are cited in Sec. 5.

D.1 Clean accuracy on D4–D6

Tab. 9 completes Tab. 1 with D4–D6. TAPQ-Net has the highest ACC in every D4–D6 condition except D5 at $\eta = 1\%$, where SAU is higher by 0.2 points under both mappings.

Table 9: **Clean accuracy (ACC, %, ↑) on D4–D6**, same protocol as Tab. 1. On D5 at $\eta = 1\%$, SAU exceeds TAPQ-Net under both mappings.

	D4			D5			D6		
Method	1	5	10	1	5	10	1	5	10
Random-class patch attack									
AAMC	77.6±0.3	75.7±0.4	74.6±0.4	73.7±0.4	73.4±0.4	72.3±0.5	71.8±0.4	71.4±0.4	70.2±0.5
QcQN	79.3±0.3	77.4±0.4	76.3±0.4	75.3±0.4	75.1±0.4	73.8±0.4	73.4±0.4	73.0±0.4	71.7±0.5
Q-Detection	80.2±0.3	78.2±0.4	77.1±0.4	76.1±0.4	75.8±0.4	74.6±0.4	74.1±0.4	73.8±0.4	72.5±0.5
QuAdRo	78.5±0.3	76.5±0.4	75.4±0.4	74.5±0.4	74.2±0.4	73.0±0.5	72.6±0.4	72.2±0.4	71.0±0.5
QUID	81.9±0.3	79.9±0.4	78.7±0.4	77.8±0.3	77.5±0.4	76.2±0.4	75.7±0.4	75.4±0.4	74.0±0.4
NQNN	81.8±0.3	81.4±0.4	81.1±0.4	80.0±0.3	79.2±0.4	77.5±0.4	77.2±0.3	76.7±0.4	75.3±0.4
TSBD	81.0±0.3	79.0±0.4	77.9±0.4	77.0±0.3	76.7±0.4	75.4±0.4	74.9±0.4	74.6±0.4	73.2±0.5
DivideMix	83.5±0.3	80.5±0.4	80.3±0.4	79.1±0.3	79.2±0.4	77.7±0.4	76.9±0.3	76.8±0.4	75.4±0.4
Co-teaching	82.7±0.3	80.7±0.4	79.5±0.4	78.6±0.3	78.3±0.4	77.0±0.4	76.5±0.4	76.1±0.4	74.8±0.4
SAU	83.1±0.3	81.3±0.4	81.1±0.4	**81.2±0.3**	79.1±0.4	78.1±0.4	77.7±0.3	76.9±0.4	76.1±0.4
NP	84.4±0.3	82.4±0.4	81.2±0.4	80.2±0.3	79.9±0.4	78.6±0.4	78.1±0.3	77.7±0.4	76.3±0.4
TAPQ-Net	**85.3±0.3**	**83.2±0.4**	**82.0±0.4**	81.0±0.3	**80.7±0.4**	**79.4±0.4**	**78.9±0.3**	**78.5±0.4**	**77.1±0.4**
All-to-one patch attack									
AAMC	76.0±0.4	74.2±0.4	73.1±0.5	72.2±0.4	71.9±0.5	70.9±0.5	70.4±0.4	70.0±0.5	68.8±0.5
QcQN	77.7±0.4	75.9±0.4	74.8±0.5	73.8±0.4	73.6±0.4	72.3±0.5	71.9±0.4	71.5±0.5	70.3±0.5
Q-Detection	78.6±0.4	76.6±0.4	75.6±0.5	74.6±0.4	74.3±0.4	73.1±0.5	72.6±0.4	72.3±0.5	71.1±0.5
QuAdRo	76.9±0.4	75.0±0.4	73.9±0.5	73.0±0.4	72.7±0.5	71.5±0.5	71.1±0.4	70.8±0.5	69.6±0.5
QUID	80.3±0.4	78.3±0.4	77.1±0.5	76.2±0.4	76.0±0.4	74.7±0.5	74.2±0.4	73.9±0.4	72.5±0.5
NQNN	80.2±0.4	79.8±0.4	79.5±0.5	78.4±0.4	77.6±0.4	76.0±0.5	75.7±0.4	75.2±0.4	73.8±0.5
TSBD	79.4±0.4	77.4±0.4	76.3±0.5	75.5±0.4	75.2±0.4	73.9±0.5	73.4±0.4	73.1±0.5	71.7±0.5
DivideMix	81.9±0.4	79.9±0.4	78.8±0.5	77.8±0.4	77.5±0.4	76.2±0.5	75.8±0.4	75.4±0.4	74.1±0.5
Co-teaching	81.0±0.4	79.1±0.4	77.9±0.5	77.0±0.4	76.7±0.4	75.5±0.5	75.0±0.4	74.6±0.4	73.3±0.5
SAU	81.4±0.4	79.7±0.4	79.5±0.5	**79.6±0.4**	77.5±0.4	76.5±0.5	76.1±0.4	75.4±0.4	74.6±0.5
NP	82.7±0.4	80.8±0.4	79.6±0.5	78.6±0.4	78.3±0.4	77.0±0.5	76.5±0.4	76.1±0.4	74.8±0.5
TAPQ-Net	**83.6±0.3**	**81.5±0.4**	**80.4±0.4**	79.4±0.4	**79.1±0.4**	**77.8±0.5**	**77.3±0.4**	**76.9±0.4**	**75.6±0.5**

D.2 Macro-F1 and AUC

Macro-F1 weights every class equally and is therefore informative for imbalanced multi-class datasets. TAPQ-Net has the highest macro-F1 in all 42 conditions (Tab. 10); under random-class poisoning it declines from 88.4 to 86.0 on OASIS and from 75.7 to 74.0 on D6 as η increases from 1% to 10%. AUC measures ranking quality independently of a decision threshold. TAPQ-Net has the highest AUC in 40 of 42 conditions (Tab. 11); NP is higher on OASIS under random-class poisoning at $\eta = 1\%$ (99.3 vs. 98.4) and $\eta = 5\%$ (98.9 vs. 98.0), and TAPQ-Net is higher at $\eta = 10\%$ (97.6 vs. 96.6).

Table 10: **Macro-F1 (%, ↑)** over three runs; columns are $\eta \in \{1, 5, 10\}\%$.

Method	OASIS			D1			D2			D3			D4			D5			D6		
	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10
Random-class patch attack																					
AAMC	80.5	80.1	78.2	78.9	78.5	76.7	80.1	79.1	78.1	75.5	74.2	73.5	74.5	72.7	71.6	70.8	70.5	69.4	68.9	68.5	67.4
QcQN	82.2	81.9	80.0	80.6	80.3	78.4	81.8	80.8	79.9	77.2	75.8	75.2	76.1	74.3	73.2	72.3	72.1	70.8	70.5	70.1	68.8
Q-Detection	83.1	82.7	80.8	81.5	81.1	79.2	82.7	81.7	80.7	77.9	76.6	76.0	77.0	75.1	74.0	73.1	72.8	71.6	71.1	70.8	69.6
QuAdRo	81.4	81.0	79.2	79.8	79.4	77.6	80.9	80.0	79.0	76.3	75.0	74.4	75.4	73.4	72.4	71.5	71.2	70.1	69.7	69.3	68.2
QUID	84.9	84.5	82.5	83.2	82.8	80.9	84.4	83.4	82.5	79.6	78.2	77.6	78.6	76.7	75.6	74.7	74.4	73.2	72.7	72.4	71.0
NQNN	84.5	84.0	82.1	82.8	82.4	80.5	84.1	83.0	82.0	79.2	77.9	77.3	78.2	76.3	75.3	74.3	74.0	72.9	72.4	72.0	70.6
TSBD	84.0	83.6	81.7	82.4	82.0	80.1	83.5	82.6	81.6	78.8	77.5	76.8	77.8	75.8	74.8	73.9	73.6	72.4	71.9	71.6	70.3
DivideMix	85.7	86.1	83.3	85.1	84.4	82.4	85.2	84.2	83.2	80.3	79.4	79.1	80.1	77.2	76.2	76.1	75.6	74.6	74.3	72.8	71.6
Co-teaching	85.8	85.4	83.4	84.1	83.7	81.8	85.3	84.3	83.3	80.5	79.1	78.4	79.4	77.5	76.3	75.5	75.2	73.9	73.4	73.1	71.8
SAU	86.6	86.2	84.3	84.9	84.5	82.6	86.2	85.1	84.2	81.3	79.9	79.3	80.2	78.3	77.2	76.2	75.9	74.7	74.3	73.9	72.5
NP	87.5	87.1	85.1	85.8	85.4	83.4	87.1	86.0	85.0	82.1	80.7	80.1	81.0	79.1	78.0	77.0	76.7	75.5	75.0	74.6	73.2
TAPQ-Net	**88.4**	**88.0**	**86.0**	**86.7**	**86.3**	**84.3**	**87.9**	**86.9**	**85.9**	**82.9**	**81.5**	**80.8**	**81.9**	**79.9**	**78.7**	**77.8**	**77.5**	**76.2**	**75.7**	**75.4**	**74.0**
All-to-one patch attack																					
AAMC	78.9	78.5	76.7	77.4	77.0	75.2	78.4	77.6	76.6	73.9	72.8	72.1	73.0	71.2	70.2	69.3	69.0	68.1	67.6	67.2	66.0
QcQN	80.6	80.2	78.4	79.0	78.6	76.9	80.2	79.2	78.2	75.6	74.3	73.6	74.6	72.9	71.8	70.9	70.7	69.4	69.0	68.6	67.5
Q-Detection	81.5	81.1	79.3	79.9	79.5	77.7	81.0	80.1	79.1	76.4	75.1	74.5	75.5	73.5	72.6	71.6	71.3	70.2	69.7	69.4	68.3
QuAdRo	79.7	79.4	77.5	78.1	77.8	76.0	79.3	78.3	77.5	74.8	73.4	72.9	73.8	72.0	70.9	70.1	69.8	68.6	68.3	68.0	66.8
QUID	83.2	82.8	80.9	81.6	81.2	79.3	82.7	81.8	80.8	78.0	76.7	76.0	77.1	75.2	74.0	73.2	73.0	71.7	71.2	70.9	69.6
NQNN	82.7	82.4	80.5	81.1	80.8	78.9	82.4	81.3	80.4	77.6	76.3	75.8	76.6	74.8	73.8	72.8	72.5	71.4	71.0	70.6	69.2
TSBD	82.3	82.0	80.0	80.7	80.4	78.4	81.9	81.0	80.0	77.3	75.9	75.3	76.2	74.3	73.3	72.5	72.2	70.9	70.5	70.2	68.8
DivideMix	84.9	84.6	82.5	83.2	82.9	80.9	84.5	83.4	82.5	79.7	78.2	77.7	78.6	76.7	75.6	74.7	74.4	73.2	72.8	72.4	71.1
Co-teaching	84.0	83.7	81.8	82.4	82.1	80.2	83.5	82.6	81.7	78.9	77.6	76.9	77.8	75.9	74.8	74.0	73.7	72.5	72.0	71.6	70.4
SAU	83.9	84.1	82.6	82.2	82.7	81.0	83.5	82.4	82.6	78.7	78.4	77.6	77.6	76.6	75.9	73.7	74.1	73.4	72.1	72.3	71.2
NP	85.8	85.4	83.3	84.1	83.7	81.7	85.4	84.3	83.3	80.5	79.1	78.5	79.4	77.5	76.4	75.5	75.2	74.0	73.5	73.1	71.7
TAPQ-Net	**86.7**	**86.3**	**84.3**	**85.0**	**84.6**	**82.6**	**86.2**	**85.2**	**84.2**	**81.3**	**79.9**	**79.2**	**80.3**	**78.2**	**77.2**	**76.2**	**75.9**	**74.7**	**74.2**	**73.8**	**72.6**

D.3 Attack success rate

Tab. 12 compares all methods in all 42 conditions; Fig. 3 in the main text shows the run-to-run variability of TAPQ-Net. TAPQ-Net has the lowest ASR in every condition. Under random-class poisoning its ASR is 0.13–0.14 on OASIS and D1, 0.14–0.16 on D2, 0.17–0.19 on D3, and 0.18–0.24 on D4–D6; under all-to-one poisoning it is 0.13–0.14 on OASIS and D1, 0.14–0.15 on D2, 0.16–0.19 on D3, 0.18–0.19 on D4, 0.22–0.23 on D5, and 0.20–0.22 on D6. The lowest-ASR competitor in every condition is AAMC (0.34–0.46), and TAPQ-Net is lower by 0.21–0.22. The three backdoor defenses, NP, SAU, and TSBD, range from 0.36 to 0.53.

D.4 Alternative frozen backbones

Tab. 13 replaces the default ResNet-18 with four frozen backbones while keeping the PCA dimension, EM branch, TQE, alignment objective, gate, and training protocol fixed. Across the four alternatives, ACC varies by at most 1.4 points per dataset, ASR by at most 0.004, and SS by at most 0.011; backbone quality therefore shifts clean utility while trigger suppression and embedding geometry stay stable.

D.5 Computational cost

Tab. 14 reports method-specific updated parameters, training time per epoch, inference latency per image, and peak memory on PathMNIST under random-class poisoning at $\eta = 10\%$. Several baselines train the full ResNet-18 (11.2M updated parameters) or two copies of it (22.4M), whereas TAPQ-Net keeps the backbone frozen. The comparison therefore holds data, trigger, and evaluation fixed but not trainable capacity. TAPQ-Net is faster per epoch than DivideMix, Co-teaching, and the four simulated-QNN baselines, and slower than AAMC, NP, SAU, and TSBD. Its per-image latency lies between the classical methods and the simulated-QNN methods.

Table 11: **AUC (%, ↑)** over three runs; columns are $\eta \in \{1, 5, 10\}\%$. NP leads on OASIS under random-class poisoning at $\eta = 1\%$ and 5%.

Method	OASIS			D1			D2			D3			D4			D5			D6		
	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10
Random-class patch attack																					
AAMC	91.4	91.0	88.8	89.6	89.2	87.1	90.9	89.8	88.7	85.7	84.3	83.6	84.6	82.5	81.3	80.3	80.0	78.8	78.3	77.8	76.5
QcQN	93.4	92.9	90.9	91.6	91.1	89.1	92.9	91.8	90.7	87.6	86.1	85.3	86.4	84.4	83.2	82.1	81.9	80.4	80.0	79.6	78.2
Q-Detection	94.4	93.9	91.7	92.5	92.1	89.9	93.8	92.8	91.7	88.5	87.0	86.3	87.4	85.2	84.0	83.0	82.6	81.3	80.8	80.4	79.0
QuAdRo	92.4	91.9	89.9	90.6	90.1	88.1	91.9	90.8	89.7	86.7	85.1	84.5	85.6	83.4	82.2	81.2	80.9	79.6	79.1	78.7	77.4
QUID	96.4	96.0	93.7	94.5	94.1	91.9	95.8	94.7	93.6	90.4	88.8	88.1	89.3	87.1	85.8	84.8	84.5	83.1	82.5	82.2	80.7
NQNN	95.9	95.5	93.2	94.0	93.6	91.4	95.4	94.3	93.2	89.9	88.5	87.7	88.8	86.7	85.4	84.3	84.1	82.7	82.1	81.7	80.3
TSBD	95.4	95.0	92.7	93.5	93.1	90.9	94.8	93.7	92.7	89.5	87.9	87.2	88.3	86.1	84.9	83.9	83.6	82.2	81.6	81.3	79.8
DivideMix	97.4	97.8	95.6	95.5	96.1	93.3	97.8	96.6	95.5	92.2	90.6	90.1	91.1	87.8	87.4	86.4	86.4	83.8	83.3	82.8	81.4
Co-teaching	97.4	96.9	94.8	95.5	95.0	92.9	96.8	95.7	94.6	91.3	89.8	89.1	90.1	87.9	86.7	85.7	85.3	83.9	83.4	83.0	81.5
SAU	98.3	97.9	95.7	96.4	96.0	93.8	97.9	96.7	95.6	92.3	90.8	90.0	91.1	88.9	87.6	86.5	86.2	84.8	84.2	83.9	82.4
NP	**99.3**	**98.9**	96.6	97.4	97.0	94.7	98.9	97.7	96.6	93.2	91.7	90.9	92.0	89.8	88.5	87.4	87.1	85.7	85.1	84.7	83.2
TAPQ-Net	98.4	98.0	**97.6**	**98.4**	**98.0**	**95.7**	**99.8**	**98.6**	**97.6**	**94.2**	**92.5**	**91.8**	**92.9**	**90.7**	**89.4**	**88.3**	**88.0**	**86.5**	**86.0**	**85.6**	**84.0**
All-to-one patch attack																					
AAMC	89.7	89.1	87.0	87.9	87.4	85.3	89.1	88.1	87.0	83.9	82.6	81.9	82.8	80.9	79.7	78.7	78.4	77.3	76.7	76.3	75.0
QcQN	91.5	91.1	89.0	89.7	89.3	87.3	91.0	89.9	88.8	85.9	84.4	83.6	84.7	82.7	81.5	80.4	80.2	78.8	78.4	77.9	76.6
Q-Detection	92.5	92.1	90.0	90.7	90.3	88.2	92.0	90.9	89.8	86.8	85.3	84.6	85.7	83.5	82.4	81.3	81.0	79.7	79.1	78.8	77.5
QuAdRo	90.5	90.1	88.0	88.7	88.3	86.3	90.0	88.9	87.9	84.9	83.4	82.8	83.8	81.8	80.6	79.6	79.2	78.0	77.5	77.2	75.9
QUID	94.6	94.0	91.8	92.7	92.2	90.0	93.8	92.9	91.8	88.5	87.1	86.3	87.5	85.3	84.0	83.1	82.8	81.4	80.9	80.6	79.0
NQNN	93.9	93.5	91.4	92.1	91.7	89.6	93.5	92.4	91.3	88.1	86.7	85.9	87.0	85.0	83.7	82.6	82.4	81.0	80.5	80.1	78.7
TSBD	93.5	93.0	90.9	91.7	91.2	89.1	93.0	91.9	90.8	87.7	86.2	85.5	86.5	84.4	83.2	82.3	82.0	80.6	80.1	79.7	78.2
DivideMix	96.4	96.0	93.7	94.5	94.1	91.9	96.0	94.7	93.6	90.5	88.8	88.2	89.3	87.1	85.9	84.8	84.5	83.1	82.6	82.2	80.8
Co-teaching	95.4	95.1	92.8	93.5	93.2	91.0	94.8	93.7	92.8	89.6	88.1	87.3	88.3	86.2	85.0	83.9	83.6	82.2	81.8	81.3	79.9
SAU	96.4	96.0	93.7	94.5	94.1	91.9	95.9	94.8	93.7	90.5	89.0	88.2	89.3	87.1	85.8	84.8	84.5	83.1	82.5	82.2	80.8
NP	97.4	97.0	94.7	95.5	95.1	92.8	96.9	95.7	94.7	91.3	89.9	89.1	90.2	88.0	86.7	85.7	85.4	84.0	83.4	83.0	81.5
TAPQ-Net	**98.4**	**97.9**	**95.6**	**96.5**	**96.0**	**93.7**	**97.9**	**96.7**	**95.6**	**92.3**	**90.7**	**90.0**	**91.1**	**88.8**	**87.6**	**86.5**	**86.2**	**84.8**	**84.3**	**83.8**	**82.4**

D.6 Reading of the ablations

Tab. 2 is read as a sequence of controlled changes. Rows (a) and (b) differ only in the encoder and show that the TQE alone does not lower ASR. Rows (a) and (c) differ only in the prototype weights $\bar{r}_i$ and give the first ASR reduction (0.36 to 0.28). Rows (d) and (e) differ only in $\mathcal{L}_{\mathrm{CL}}$ with a prototype state that is fixed after initialization, and alignment to that fixed state does not lower ASR. Rows (e) and (f) differ only in the gate of Eq. (6), and this change produces the largest improvement on every metric at an unchanged parameter count. Because refresh is the only mechanism through which the EM state evolves after initialization, rows (e) and (f) also differ in the number of EM passes; App. G.2 specifies the controls that separate the two effects. In Tab. 3, all encoders are embedded in the complete loop, so the comparison isolates the encoder but not its interaction with the gate.

D.7 Paired significance analysis

We test whether the advantage of TAPQ-Net over its strongest comparator is systematic across the evaluation grid. We pair TAPQ-Net with NP, which has the strongest condition-level results among the compared methods, and treat each of the 42 dataset–mapping–ratio conditions as one paired unit; the three runs within a condition are averaged first and never counted as independent observations. For each metric we report the mean paired difference, a 95% bootstrap confidence interval over conditions (20,000 resamples), the two-sided Wilcoxon signed-rank test with Holm correction across the four metrics, the matched-pairs rank-biserial correlation, and the sign counts. We compute all quantities directly from the condition-level means in Tabs. 1–12, so any reader can reproduce them from the reported values.

Tab. 15 shows that the ACC, ASR, and macro-F1 differences have the same sign in every condition, which is why their rank-biserial correlations reach ± 1; the two unfavorable AUC conditions are the OASIS random-class entries at $\eta \in \{1, 5\}\%$. Because conditions share datasets and methods, the test establishes consistency across the evaluation grid rather than independence of the 42 observations.

Table 12: **ASR (↓)** over three runs; columns are $\eta \in \{1, 5, 10\}\%$. ASR definitions are given in App. H.3.

Method	OASIS			D1			D2			D3			D4			D5			D6		
	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10
Random-class patch attack																					
AAMC	0.34	0.35	0.35	0.35	0.36	0.36	0.36	0.37	0.38	0.39	0.39	0.41	0.40	0.41	0.42	0.44	0.45	0.46	0.43	0.44	0.45
QcQN	0.35	0.36	0.36	0.36	0.37	0.37	0.37	0.38	0.39	0.40	0.40	0.42	0.41	0.42	0.43	0.45	0.46	0.47	0.44	0.45	0.46
Q-Detection	0.35	0.36	0.37	0.36	0.37	0.38	0.38	0.39	0.39	0.40	0.41	0.43	0.42	0.42	0.44	0.46	0.47	0.47	0.44	0.46	0.47
QuAdRo	0.34	0.35	0.35	0.35	0.36	0.36	0.36	0.37	0.38	0.39	0.39	0.41	0.40	0.41	0.42	0.44	0.45	0.46	0.43	0.45	0.45
QUID	0.36	0.37	0.38	0.37	0.38	0.39	0.39	0.40	0.40	0.41	0.42	0.44	0.43	0.43	0.45	0.47	0.48	0.48	0.45	0.47	0.48
NQNN	0.40	0.41	0.41	0.41	0.42	0.42	0.43	0.44	0.44	0.45	0.45	0.47	0.46	0.47	0.48	0.50	0.52	0.53	0.49	0.52	0.52
TSBD	0.36	0.37	0.37	0.37	0.38	0.38	0.38	0.40	0.40	0.41	0.42	0.43	0.42	0.43	0.44	0.46	0.47	0.48	0.45	0.47	0.47
DivideMix	0.38	0.39	0.40	0.39	0.40	0.41	0.41	0.42	0.42	0.43	0.44	0.45	0.45	0.45	0.46	0.49	0.49	0.50	0.47	0.49	0.49
Co-teaching	0.37	0.38	0.38	0.38	0.39	0.39	0.40	0.41	0.41	0.42	0.43	0.45	0.44	0.44	0.45	0.47	0.48	0.49	0.46	0.48	0.48
SAU	0.39	0.40	0.40	0.40	0.41	0.41	0.42	0.43	0.43	0.44	0.44	0.46	0.45	0.46	0.47	0.49	0.51	0.52	0.48	0.51	0.51
NP	0.39	0.40	0.40	0.40	0.41	0.41	0.42	0.43	0.43	0.44	0.44	0.46	0.45	0.46	0.47	0.49	0.50	0.51	0.48	0.50	0.50
TAPQ-Net	**0.13**	**0.14**	**0.14**	**0.13**	**0.14**	**0.14**	**0.14**	**0.15**	**0.16**	**0.17**	**0.17**	**0.19**	**0.18**	**0.19**	**0.20**	**0.22**	**0.23**	**0.24**	**0.21**	**0.22**	**0.23**
All-to-one patch attack																					
AAMC	0.34	0.34	0.35	0.35	0.35	0.36	0.36	0.37	0.37	0.38	0.39	0.41	0.40	0.40	0.41	0.44	0.44	0.45	0.42	0.44	0.44
QcQN	0.35	0.35	0.36	0.36	0.36	0.37	0.37	0.38	0.38	0.39	0.40	0.42	0.41	0.41	0.43	0.45	0.45	0.46	0.43	0.45	0.45
Q-Detection	0.35	0.36	0.36	0.36	0.37	0.37	0.38	0.39	0.39	0.40	0.41	0.42	0.42	0.42	0.43	0.45	0.46	0.47	0.44	0.46	0.46
QuAdRo	0.34	0.35	0.35	0.35	0.36	0.36	0.36	0.37	0.38	0.39	0.39	0.41	0.40	0.41	0.42	0.44	0.45	0.46	0.43	0.44	0.45
QUID	0.36	0.37	0.37	0.37	0.38	0.38	0.39	0.40	0.40	0.41	0.42	0.43	0.43	0.43	0.44	0.46	0.47	0.48	0.45	0.47	0.47
NQNN	0.41	0.42	0.42	0.42	0.43	0.43	0.44	0.45	0.45	0.46	0.46	0.48	0.47	0.48	0.49	0.51	0.53	0.54	0.50	0.53	0.53
TSBD	0.36	0.37	0.37	0.37	0.38	0.38	0.38	0.39	0.39	0.40	0.41	0.43	0.42	0.42	0.44	0.46	0.47	0.47	0.44	0.46	0.47
DivideMix	0.38	0.39	0.39	0.39	0.40	0.40	0.40	0.41	0.42	0.43	0.43	0.45	0.44	0.45	0.46	0.48	0.49	0.50	0.47	0.48	0.49
Co-teaching	0.37	0.38	0.38	0.38	0.39	0.39	0.40	0.41	0.41	0.42	0.43	0.44	0.43	0.44	0.45	0.47	0.48	0.49	0.46	0.48	0.48
SAU	0.40	0.41	0.41	0.41	0.42	0.42	0.43	0.44	0.44	0.45	0.45	0.47	0.46	0.47	0.48	0.50	0.52	0.53	0.49	0.52	0.52
NP	0.40	0.41	0.41	0.41	0.42	0.42	0.43	0.44	0.44	0.45	0.45	0.47	0.46	0.47	0.48	0.50	0.51	0.52	0.49	0.51	0.51
TAPQ-Net	**0.13**	**0.13**	**0.14**	**0.13**	**0.13**	**0.14**	**0.14**	**0.15**	**0.15**	**0.16**	**0.17**	**0.19**	**0.18**	**0.18**	**0.19**	**0.22**	**0.22**	**0.23**	**0.20**	**0.22**	**0.22**

Table 13: **Alternative frozen backbones**, random-class, $\eta = 10\%$. The default ResNet-18 corresponds to the $\eta = 10\%$ random-class entries of Tabs. 1, 9, and 12. ASR and SS are shown without the leading zero.

Backbone	OASIS			D1			D2			D3			D4			D5			D6		
	ACC	ASR	SS	ACC	ASR	SS	ACC	ASR	SS	ACC	ASR	SS	ACC	ASR	SS	ACC	ASR	SS	ACC	ASR	SS
DenseNet-121	89.2	.159	.687	87.4	.139	.657	89.1	.159	.667	83.8	.189	.607	81.6	.199	.587	79.0	.229	.557	76.7	.239	.537
EfficientNet-B0	88.9	.159	.685	87.1	.139	.655	88.8	.159	.665	83.6	.189	.605	81.4	.199	.586	78.8	.228	.556	76.5	.238	.536
ConvNeXt-Tiny	88.3	.158	.680	86.5	.138	.650	88.2	.158	.660	82.9	.187	.601	80.8	.197	.581	78.2	.227	.552	75.9	.236	.532
ViT-B/16	87.8	.157	.676	86.1	.137	.647	87.7	.157	.657	82.5	.186	.598	80.4	.196	.578	77.8	.225	.549	75.6	.235	.529

Table 14: **Computational cost** on PathMNIST, random-class, $\eta = 10\%$. All methods use the same preprocessing, batch size, hardware, and poisoned split; frozen feature extractors are excluded from updated-parameter counts.

Method	Category	Updated params	Time/epoch (s)	Latency/img (ms)	Peak mem. (GB)	Inference path
AAMC	Medical adversarial reference	11,181,642	23.8	1.08	2.10	CNN classifier
Co-teaching	Noisy-label learning	22,363,284	48.6	1.12	4.05	Single selected network
DivideMix	Noisy-label learning	22,363,284	72.4	1.35	4.72	Single selected network
NP	Backdoor defense	66,048	17.9	1.24	2.38	Backdoored model + polarizer
NQNN	Hybrid quantum medical	84,392	54.8	3.75	2.91	Encoder + VQC + classifier
QcQN	Quantum certified robustness	90,240	63.2	4.10	3.02	Quantum-noise classifier
Q-Detection	Quantum poisoning detection	2,049	58.6	1.19	3.58	Filtered-data classifier
QuAdRo	Quantum certified robustness	92,480	118.7	8.80	3.21	Smoothed quantum classifier
QUID	QNN data-poisoning attack	88,736	60.4	4.35	2.84	QNN under poisoning protocol
SAU	Backdoor defense	11,181,642	28.9	1.16	2.64	Purified classifier
TSBD	Backdoor defense	11,181,642	34.7	1.22	2.81	Reinitialized + fine-tuned classifier
TAPQ-Net	Prototype–quantum defense	**178**	39.6	2.65	2.53	Encoder + PCA + TQE + classifier

D.8 TOPOLOGY-MATCHED CLASSICAL TREE

The lower block of Tab. 3 replaces the TQE with a classical encoder that keeps its interaction graph and budget: the same pairwise ($\mathcal{P}_1$), branch ($\mathcal{P}_2$), and root ($\mathcal{P}_3$) aggregation, the same eight outputs read by the unchanged 81-parameter classifier, and therefore the same 88 gradient-trained parameters. Each of its seven internal nodes combines the outputs of its two children through one trainable parameter, which reproduces the seven-parameter budget of the TQE. The EM branch, trust

Table 15: **Paired comparison of TAPQ-Net with NP over the 42 conditions.** Differences are TAPQ-Net minus NP; for ASR, negative favors TAPQ-Net. Signs count conditions that favor / disfavor TAPQ-Net.

Metric	Mean diff.	95% CI	Holm p	Rank-biserial	Signs (+/−)
ACC (points)	$+0.85$	$[+0.83, +0.87]$	2.8×10^{-8}	$+1.00$	42 / 0
ASR	-0.277	$[-0.280, -0.275]$	3.0×10^{-8}	-1.00	42 / 0
Macro-F1 (points)	$+0.82$	$[+0.80, +0.85]$	3.0×10^{-8}	$+1.00$	42 / 0
AUC (points)	$+0.82$	$[+0.69, +0.91]$	1.6×10^{-7}	$+0.93$	40 / 2

weighting, alignment objective, gate, optimizer, splits, and poison realizations are identical, and both encoders are rerun with three new seeds. The TQE reaches 88.59 ± 0.18 ACC, 0.157 ± 0.003 ASR, and 0.664 ± 0.003 SS, against 85.21 ± 0.24, 0.191 ± 0.004, and 0.633 ± 0.005 for the classical tree. The rerun TQE value lies 0.91 ACC points below the full-model entry of Tab. 2(f), which reflects seed-to-seed variation between independent runs; we compare the two encoders only within the matched rerun.

D.9 Trust-score diagnostics

Tab. 4a groups the final trust score r_i on the 89,996-sample PathMNIST training split, of which 80,996 samples are clean and 9,000 are poisoned. We consult the poison annotations only after training; they never enter EM fitting, optimization, or model selection. The mean trust score of poisoned samples is 0.0972 lower than that of clean samples, and the AUROC of r_i for ranking clean above poisoned samples is 0.724. Two consequences follow from Eqs. (3)–(5). First, the mean cross-entropy weight of poisoned samples is 12.3% lower (ratio $0.694/0.791$), which reduces the influence of their labels without any hard decision. Second, poisoned targets keep a larger share of their own latent observation (mean $1 - r_i$ of 0.306 against 0.209), which is the residual term of Corollary 4.2. The distributions overlap substantially, so we do not treat r_i as a poison detector.

D.10 PCA dimensionality

Tab. 4b retains $d_z \in \{2, 4, 6, 8\}$ principal components of the 512-dimensional training-split features and reports the cumulative explained variance. For $d_z < 8$ we zero-pad z_i^c to eight coordinates; since $R_Y(0)|0\rangle = |0\rangle$, the padded qubits enter the TQE in $|0\rangle$, and the circuit, classifier, and training pipeline stay unchanged. ACC, ASR, and SS improve monotonically with d_z, and the gain shrinks from 5.33 ACC points ($2 \to 4$) and 4.51 ($4 \to 6$) to 1.16 ($6 \to 8$). Dimensions above eight would require additional qubits, grouped encoding, or data re-uploading and therefore change more than one variable at a time.

D.11 Clean-label and warping attacks

Tab. 5b evaluates two triggers that differ from the visible patch in kind rather than in size or location. SIG (Barni et al., 2019) superimposes a sinusoidal signal on a fraction of target-class training images without changing any label; WaNet (Nguyen & Tran, 2021) applies a smooth spatial warping field and maps triggered samples to one target class. Both run on PathMNIST at a 10% poisoning rate with three runs, and NQNN serves as the comparator. Both attacks follow their original formulations (Barni et al., 2019; Nguyen & Tran, 2021). Relative to NQNN, TAPQ-Net lowers ASR by 0.120 under SIG and by 0.129 under WaNet and raises ACC by 2.20 and 2.08 points.

E Positioning Relative to the Closest Prior Work

Tab. 16 compares TAPQ-Net with the methods closest to it in mechanism. Several ingredients are shared with prior work: frozen features, PCA, and a Gaussian mixture appear in Kutbi (2025); prototypes are used against backdoors by Amula et al. (2025) and Guo et al. (2025b); sample reweighting on poisoned data appears in Zhang et al. (2023); and prototype regularization of a quantum medical model appears in Rahman & Zhuang (2026b). The element that none of these methods contains is the reverse channel, in which the geometry of the representation being trained determines whether the supervisory geometry may be updated.

Table 16: **Mechanism-level comparison** with the closest prior work.

Method	Stage	Requires	Geometry and its update	Effect on training
Activation clustering (Chen et al., 2019)	detect, retrain	model trained on $\mathcal{D}^p$	per-class activation clusters, estimated once	flagged samples removed
Kutbi (2025)	filter, retrain	trusted pretrained extractor	PCA and GMM on frozen features, estimated once	flagged samples removed
ACDM (Zhang, 2026)	detect, retrain	QNN trained on $\mathcal{D}^p$	PCA and two-cluster k-means on measurement activations, estimated once	flagged samples removed
PGBD (Amula et al., 2025)	post hoc	backdoored model	activation-space prototypes for a fine-tuning stage	sanitization loss penalizing displacement toward the trigger direction
CBD (Zhang et al., 2023)	in training	auxiliary backdoored model	embeddings of the auxiliary model	sample reweighting and mutual-information minimization
PGRL (Guo et al., 2025b)	in training	verified-benign subset	prototypes from the benign subset	label-consistency filtering and feature-distance steering
QAPNet (Rahman & Zhuang, 2026b)	in training (input corruption)	–	prototype-contrastive regularization with quantum encoding	attention-based instance reweighting
TAPQ-Net	in training	frozen encoder only	label-free EM mixture on fixed PCA latent; one EM pass per epoch, only if the quantum-embedding gate passes	continuous alignment targets s_i and weights $\bar{r}_i$; no sample removed

(i) PGBD. PGBD sanitizes a model that has already learned the backdoor, using a post-hoc fine-tuning stage. TAPQ-Net never trains a backdoored reference: the prototypes are estimated before and during training, are label-free, and enter as sample-specific convex targets s_i rather than as class prototypes of a backdoored network.

(ii) PCA and GMM filtering. Kutbi (2025) and TAPQ-Net share the front end. The former converts the mixture into a single keep/remove decision and retrains on the retained set. TAPQ-Net retains every sample, uses responsibilities continuously in the objective, and re-estimates the mixture only when the gate passes; an assignment made at one refresh can therefore be revised at a later accepted refresh, whereas a removed sample cannot be recovered.

(iii) QAPNet. QAPNet targets input corruption and bag-level annotation, and uses prototype-contrastive regularization alongside quantum encoding and attention reweighting. TAPQ-Net targets training-time backdoors, estimates prototypes without labels, and conditions their update on the geometry of the quantum embedding.

(iv) QNN-specific backdoor defenses. ACDM (Zhang, 2026) and QSentry (Wang et al., 2025) cluster the measurement outputs of a QNN trained on poisoned data in order to detect and remove poisoned samples. TAPQ-Net uses quantum measurement geometry not to classify samples as poisoned but to decide whether the prototype state may evolve.

The incremental value of the reverse channel is supported by the change from row (e) to row (f) of Tab. 2, subject to the caveat of App. D.6, and the contribution of the quantum realization beyond its tree topology by App. D.8.

F Baseline Protocol

All methods are trained on the same poisoned training split with identical poisoned indices, trigger operator, preprocessing, and clean test set, and are evaluated with the metrics of App. H.3. Methods that were not designed as backdoor defenses are trained on the poisoned split with their original objective and evaluated as reference points. Tab. 17 states, for each method, the problem it was designed for and how we apply it.

G Evaluation Boundary and Specified Controls

This section separates the controls that our experiments now answer from those that remain open, and specifies each open control precisely enough to be run. We assume none of the open outcomes in the paper.

Table 17: **Compared methods**: the problem each method was designed for and its application in our benchmark. All methods run in our pipeline under the shared protocol of App. H.3.

Method	Designed for	Applied here as
NP	backdoor purification	polarizer layer trained on the model obtained from $\mathcal{D}^p$
SAU	backdoor unlearning	shared adversarial unlearning of the model obtained from $\mathcal{D}^p$
TSBD	two-stage backdoor defense	reinitialization and fine-tuning of the model obtained from $\mathcal{D}^p$
DivideMix	noisy-label learning	co-divide and semi-supervised training on $\mathcal{D}^p$
Co-teaching	noisy-label learning	two networks with small-loss sample exchange on $\mathcal{D}^p$
AAMC	adversarial attacks on medical classifiers	CNN classifier trained on $\mathcal{D}^p$; reference
QcQN	certified robustness via quantum noise	quantum-noise classifier trained on $\mathcal{D}^p$
QuAdRo	randomized-smoothing certification	quantum classifier trained on $\mathcal{D}^p$ with smoothed inference
Q-Detection	quantum poisoning detection	detection on $\mathcal{D}^p$ followed by training on the retained set
QUID	QNN data-poisoning attack	QNN trained on $\mathcal{D}^p$ under its protocol; reference, not a defense
NQNN	noise-aware medical QNN	encoder, VQC, and classifier trained on $\mathcal{D}^p$

G.1 Answered controls

Four controls that bear directly on our claims are answered on PathMNIST at $\eta = 10\%$. *Topology versus realization*: a classical tree with the TQE's graph and budget trails the TQE by 3.38 ACC points (App. D.8). *Trust semantics*: poisoned samples receive lower trust scores (AUROC 0.724), and the overlap confirms that r_i interpolates rather than detects (App. D.9). *Compression*: eight components retain 93.2% of the feature variance, and the returns of additional components diminish (App. D.10). *Statistics*: paired tests over the 42 conditions are significant for all four metrics (App. D.7). Extending the first three studies to all seven datasets and to the all-to-one mapping is the natural next step.

G.2 Refresh controls

To separate gating from continued EM refinement, the following variants should share every other setting of row (f) of Tab. 2: always-refresh (every Δ_{ref} epochs, no gate), silhouette-only and balance-only gates, a gate with shuffled assignments $\{a_i\}$, and a fixed state obtained by running EM to convergence at initialization. Reporting the number of accepted refreshes per run belongs to this study.

G.3 Adaptive and additional attacks

The main evaluation uses visible patch triggers with random-class and all-to-one mappings at 1%, 5%, and 10%; Tab. 5a varies patch size, location, and opacity; and Tab. 5b adds the clean-label SIG and the warping-based WaNet attacks on one dataset against one comparator. Frequency-domain and scanner- or site-artifact triggers remain untested. The most informative adaptive attack targets the failure mode of Sec. 4: the attacker optimizes the trigger so that the latent offsets δ_i make triggered samples form a compact, well-separated component, maximizing the silhouette and balance statistics of Eq. (6) while maximizing ASR, in the spirit of attacks that manipulate latent separability (Qi et al., 2023). The trust diagnostics of Tab. 4a provide the natural monitor for this attack: a successful adaptive trigger would raise the trust scores of poisoned samples toward that of clean samples.

G.4 Execution on quantum hardware

We simulate the circuit and do not model device noise. Proposition A.3 shows that equal-strength depolarizing and symmetric readout noise leave the alignment loss and the gate unchanged; gate-level noise, amplitude damping, and finite-shot estimation are not covered and require simulation with a noise model or execution on hardware.

G.5 Clinical reliability

The results are benchmark evidence. Clinical use would additionally require external multi-site evaluation, calibration (Brier score and negative log-likelihood), risk–coverage analysis, robustness to scanner and acquisition shift, subgroup assessment, and prospective validation. The protocol also assumes a clean validation split for epoch selection; where validation data may be poisoned, a label-free selection rule would be required.

H Implementation and Reproducibility

H.1 Datasets

Tab. 18 summarizes the seven datasets. For OASIS, subjects are partitioned by patient identifier before slice extraction and poisoning, so all slices of a patient lie in a single partition, and poisoning is applied only to slices of the training partition.

Table 18: **Datasets.** D1–D6 identify the MedMNISTv2 datasets throughout the paper.

ID	Dataset	Scale / classes	Task
OASIS	OASIS MRI	80,000+ axial slices, 461 patients, 4 CDR stages	dementia-stage classification
D1	BloodMNIST	17,092 images, 8 classes	blood-cell classification
D2	PathMNIST	107,180 images, 9 classes	colorectal histology classification
D3	OrganAMNIST	58,218 images, 11 classes	abdominal CT organ classification (axial)
D4	OrganCMNIST	24,292 images, 11 classes	abdominal CT organ classification (coronal)
D5	OCTMNIST	109,309 images, 4 classes	retinal OCT disease classification
D6	DermaMNIST	10,015 images, 7 classes	dermatoscopic lesion classification

H.2 Configuration

Tab. 19 gives the default configuration, which is shared across datasets and label mappings; the hyperparameters of the EM branch and the gate are in Tab. 6.

Table 19: **Default TAPQ-Net configuration.**

Component	Setting
Frozen feature extractor	ResNet-18 (DenseNet-121, EfficientNet-B0, ConvNeXt-Tiny, ViT-B/16 in Tab. 13)
Feature / latent dimension	512 / $d_z = 8$
Latent projection	PCA fitted once on the poisoned training split, then fixed
Prototype branch	label-free EM, $K = C$ isotropic components, EMA-damped passes
Encoder / readout	8-qubit TQE with 7 trainable rotations / Pauli-Z expectations
Classifier	linear, $W_{\mathrm{cls}} \in \mathbb{R}^{C\times 8}$, $b \in \mathbb{R}^C$
Objective	$\mathcal{L}_{\mathrm{CE}}$ with weights $\bar{r}_i$ plus $\lambda_{\mathrm{CL}}\mathcal{L}_{\mathrm{CL}}$ (Eq. (5))
Refresh	gate of Eq. (7)
Optimizer / learning rate / batch size	Adam / 10^{-3} / 128
Training budget	at most 15 epochs, epoch selected on the clean validation split
Metrics / runs	ACC, macro-F1, AUC, ASR, SS / three independent runs

H.3 Attack protocol and metrics

For every method, a fraction $\eta \in \{1, 5, 10\}\%$ of training images is selected under the run seed and stamped with a 3×3 white square in the bottom-right corner before normalization. Random-class poisoning assigns each selected sample a label drawn from $\mathcal{Y} \setminus \{y_i\}$; all-to-one poisoning assigns the fixed target t. Validation and test images are unmodified. With $\mathcal{D}_{\mathrm{te}}$ the clean test set and $\widehat{y}(\cdot)$ the

prediction,

$$\mathrm{ASR}_{\mathrm{a2o}} = \frac{\big|\{x \in \mathcal{D}_{\mathrm{te}} : y(x) \neq t,\ \widehat{y}(T(x)) = t\}\big|}{\big|\{x \in \mathcal{D}_{\mathrm{te}} : y(x) \neq t\}\big|},$$

$$\mathrm{ASR}_{\mathrm{rand}} = \frac{1}{|\mathcal{D}_{\mathrm{te}}|} \sum_{x \in \mathcal{D}_{\mathrm{te}}} \mathbf{1}\big[\widehat{y}(T(x)) = t(x)\big],$$

where $t(x) \in \mathcal{Y} \setminus \{y(x)\}$ is an incorrect label fixed by the evaluation seed. The random-class ASR therefore measures agreement with a seed-fixed incorrect mapping, not arbitrary misclassification. ACC, macro-F1, and AUC are computed on clean test images. SS is the silhouette coefficient of the learned quantum embeddings $\{z_i^q\}$ and measures their compactness and separation.

H.4 Software and hardware

Experiments use Python 3.12.12, PyTorch 2.8.0+cu126, and PennyLane 0.44.0 on two NVIDIA Tesla T4 GPUs (14.56 GB each). We simulate the quantum circuit in PennyLane and do not model device noise. PCA is never fitted on validation or test data, and frozen feature extractors are excluded from all parameter counts.

I Extended Related Work

Hybrid quantum–classical medical image classification. Hybrid learning has entered medical image analysis through quantum convolutional structures, quantum transfer learning, and compact variational heads (Landman et al., 2022; Bokhan et al., 2022; Dhara et al., 2024; Ajlouni et al., 2023; Yousif et al., 2024), which act as parameter-efficient nonlinear modules when data are limited. Noise-aware medical quantum models adapt the quantum branch to corrupted supervision (NQNN) or corrupted inputs (QAPNet) (Rahman & Zhuang, 2026a;b). These studies evaluate clean supervision, label noise, or input corruption; TAPQ-Net instead treats the quantum branch as a prototype-supervised embedding generator under training-time backdoor poisoning.

Backdoor poisoning and defense. Backdoor attacks implant trigger-dependent behavior while preserving clean accuracy; beyond image-and-label poisoning, recent work demonstrates label-only and multimodal variants (Gu et al., 2017; Jha et al., 2023; Yang et al., 2023b), clean-label attacks that leave every label intact (Turner et al., 2019; Barni et al., 2019), blended triggers (Chen et al., 2017), and imperceptible warping triggers (Nguyen & Tran, 2021). Defenses have evolved from sample filtering (Tran et al., 2018; Chen et al., 2019) to feature purification, adversarial unlearning, and input purification: Neural Polarizer filters trigger-related features through an inserted layer (Zhu et al., 2023), SAU unlearns shared adversarial examples between poisoned and purified models (Wei et al., 2023), TSBD reinitializes and fine-tunes (Lin et al., 2024), and ZIP and SampDetox restore poisoned inputs by transformation or diffusion (Shi et al., 2023; Yang et al., 2024). Because low ASR after purification may leave backdoor-connected features intact (Min et al., 2024), TAPQ-Net shapes the representation during training rather than repairing it afterwards. Adaptive attacks that suppress latent separability (Qi et al., 2023) define the strongest test of any geometry-based defense, including ours (App. G.3).

Quantum robustness and poisoning-aware quantum learning. Quantum robustness has been studied through quantum noise, randomized smoothing, certified robustness, and poisoning detection or analysis (Du et al., 2021; Huang et al., 2022a; Sahdev & Kumar, 2023; He et al., 2025; Kundu & Ghosh, 2025), and QNN-specific backdoor defenses detect poisoned samples by clustering measurement outputs (Zhang, 2026; Wang et al., 2025). These methods certify, detect, or remove; TAPQ-Net uses measurement geometry to decide whether its prototype targets may evolve.

Prototype and contrastive representation learning. Prototypical networks show that class prototypes provide a strong inductive bias in low-data classification (Snell et al., 2017), and prototypical and EM-style contrastive learning show that latent prototypes improve semantic compactness beyond instance-level objectives (Li et al., 2021a; Jin et al., 2022). Hard labels and fixed centroids are unreliable when the training distribution is poisoned, so TAPQ-Net uses soft, sample-specific targets that interpolate between the mixture mean and the observation according to trust score (Eq. (3)).

Tree-structured variational quantum encoders. Tree tensor networks factorize high-dimensional structure into local contractions and aggregate information progressively across a tree (Orús, 2014; Stoudenmire & Schwab, 2016; Huggins et al., 2019); hierarchical quantum classifiers and tensor-network-inspired circuits carry this structure into parameter-efficient quantum learning (Grant et al., 2018; Huggins et al., 2019; Cong et al., 2019), and QCNN-style hierarchies can mitigate barren plateaus relative to unconstrained ansätze (McClean et al., 2018; Cerezo et al., 2021; Pesah et al., 2021). Our contribution is not a new circuit: we use the TQE as a constrained embedding generator and make it poisoning-aware by coupling it with label-free soft prototypes, trust weighting, prototype–quantum contrastive alignment, and feedback-controlled prototype refresh.