QMORE: ROBUST QUANTUM IMAGE–TEXT RETRIEVAL VIA ENCODER SHARING AND DIVERSITY-AWARE RANK FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present QMORE, a hybrid quantum–classical framework for cross-modal image–text retrieval. QMORE learns a shared image–text geometry by reusing a single variational quantum encoder across both modalities, paired with lightweight modality adapters that preserve modality-specific flexibility while keeping quantum parameters shared. To reduce run-to-run sensitivity, QMORE ensembles independently trained models, weights members by retrieval strength and diversity, and fuses predictions via weighted rank voting. Across three benchmarks, QMORE improves Recall@1 over the strongest parameter-matched baseline by $+3.2$, $+2.2$, and $+2.8$ points on SVO Probes, Captioned CIFAR-10, and Flickr30k—a 24–26% relative improvement on every dataset—while also achieving the best mean and median rank throughout. The advantage persists against classical encoders with $5\times$ more encoder-core parameters, and under a calibrated noise model of a near-term IBM device QMORE retains 92% of its Recall@1 while the quantum baseline collapses to near-random.

1 INTRODUCTION

Quantum machine learning (QML) is attracting interest as a representation learning paradigm that encodes data into quantum states, exploiting superposition and entanglement to induce expressive feature maps in high-dimensional Hilbert spaces with compact parameterizations Havlíček et al. (2019); Schuld & Killoran (2019). This perspective is especially relevant as improving classical performance increasingly comes from scaling parameters, compute, and training data, motivating interest in whether quantum-induced representations can improve expressivity per parameter beyond what scaling alone provides Huang et al. (2021).

At the same time, image–text retrieval has become increasingly central to vision–language systems, serving both as a practical capability for cross-modal search and as a widely used benchmark for evaluating multimodal representation quality Radford et al. (2021); Jia et al. (2021); Zhou et al. (2022); Koh et al. (2023). Given an image query, the task is to retrieve the most relevant caption by embedding both modalities into a shared space. In large-scale settings, contrastive pretraining with classical neural architectures (e.g., CLIP-style models) has proven highly effective Radford et al. (2021); Jia et al. (2021). Yet at its core, multimodal retrieval is the task of learning a *shared geometry*: a metric space in which semantic relationships are preserved across modalities. Quantum and hybrid quantum–classical models are a natural fit for capturing such richly coupled structure, as embeddings are encoded as states in a high-dimensional Hilbert space where superposition represents structured combinations of basis configurations and entanglement enables non-separable feature correlations.

However, deploying QML for image–text retrieval raises two practical challenges: **(i) parameter replication across modalities** and **(ii) run-to-run variability in variational training**. Typically, quantum multimodal pipelines allocate *separate* quantum encoders to each modality Chen et al. (2025). But duplicating circuits inflates the variational parameter footprint and undermines parameter efficiency—an important consideration in quantum settings, where optimization is sensitive to how variational capacity is allocated Larocca et al. (2025). Moreover, variational quantum models often exhibit substantial run-to-run variability: performance can depend strongly on initialization

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.

and circuit design, and increasing expressivity can worsen trainability via vanishing gradients and related phenomena McClean et al. (2018); Larocca et al. (2025). Together, these constraints motivate QMORE's design: a quantum retrieval architecture that (1) enforces a coherent shared embedding geometry *without duplicating* quantum parameters, and (2) improves robustness by converting training variability into complementary retrieval signals.

We realize this design with QMORE, a hybrid quantum–classical embedding framework that reuses a single shared quantum encoder across modalities via lightweight adapters, and stabilizes retrieval with an accuracy–diversity weighted rank-fusion ensemble. Our key contributions include:

- **Quantum multimodal retrieval.** To our knowledge, we present the first hybrid quantum–classical pipeline for image–text retrieval, including training, retrieval evaluation, and analysis.
- **Shared encoder with modality adapters.** We introduce a parameter-efficient quantum multimodal architecture that reuses one variational encoder across modalities, with lightweight modality-specific adapters.
- **Accuracy–diversity weighted rank ensemble.** Motivated by run-to-run variability in variational training, we propose a novel ensembling mechanism that prioritizes accurate and complementary members using embedding-level diversity, and aggregates predictions via weighted rank voting. Though designed for quantum models, it also improves classical retrieval ensembles.
- **NISQ-era noise robustness.** We evaluate QMORE on Flickr30k under a calibrated noise model of IBM's `ibmq_guadalupe` device, incorporating calibrated gate, thermal-relaxation, and readout errors. QMORE retains 92% of its Recall@1, while the quantum baseline degrades to near-random performance, suggesting QMORE remains practical in the NISQ era.

2 Background and Related Work

Image–text retrieval and shared representation learning. Image–text retrieval aims to learn a shared space where semantically matched images and captions are close and mismatched pairs are separated. Contrastive objectives such as InfoNCE increase the similarity of matched pairs relative to other examples in the batch Oord et al. (2018); Radford et al. (2021); Jia et al. (2021).

Early work established joint image–sentence embedding and retrieval on paired vision–language benchmarks Hodosh et al. (2013); Karpathy & Fei-Fei (2015); Plummer et al. (2015), with later methods improving cross-modal alignment Faghri et al. (2018); Lee et al. (2018). Modern systems such as CLIP and ALIGN scale this paradigm through web-scale contrastive pretraining and achieve strong image–text retrieval performance Radford et al. (2021); Jia et al. (2021). These systems provide important practical reference points, but their large architectures and extensive paired pretraining make them fundamentally different from the low-parameter regime studied here.

Prior work shows that modality-shared contrastive pretraining can preserve retrieval quality while reducing parameter duplication You et al. (2022). This motivates mechanisms that produce expressive shared embedding spaces with limited parameters and suggests that, under a constrained budget, capacity should be concentrated in a common representation rather than replicated across modality-specific encoders.

Quantum representations under constrained budgets. Quantum feature maps embed classical data into states in a high-dimensional Hilbert space, while variational quantum circuits introduce trainable transformations through a comparatively small number of parameterized gates Havlíček et al. (2019); Schuld & Killoran (2019). Entangling operations can further couple input features through non-separable transformations, providing a distinct inductive bias for learning structured representation geometries.

Compact quantum parameterization does not by itself imply quantum advantage: a large Hilbert space guarantees neither trainability nor superiority over classical alternatives McClean et al. (2018). We therefore pose a precise empirical question: under controlled, parameter-matched conditions, do variational quantum embeddings provide an effective inductive bias for multimodal alignment?

Answering this question requires considering not only representation quality, but also whether that inductive bias can be realized under the constraints of near-term variational models. Available circuits are limited in qubit count and depth, while increasing circuit complexity can amplify noise exposure and exacerbate optimization difficulties, including vanishing gradients Larocca et al. (2025); McClean et al. (2018). These constraints make the allocation of variational capacity a central architectural consideration, particularly when multiple modalities must be represented in a shared space.

Quantum multimodal representation learning. Existing quantum multimodal work has primarily considered alignment or prediction tasks outside image–text retrieval. Chen et al. align EEG and image features using modality-specific processing pipelines and separate variational quantum circuits Chen et al. (2025). Other hybrid approaches use variational quantum layers to fuse heterogeneous representations for downstream tasks such as multimodal sentiment analysis Zheng et al. (2025).

A direct adaptation of modality-specific quantum pipelines to image–text retrieval would assign a separate variational encoder to each modality. Under a constrained quantum parameter budget, this duplicates circuit parameters across the image and text pathways, dividing limited variational capacity between two independently learned mappings that contrastive training must then reconcile. A shared circuit instead applies the same learned transformation to both modalities, directly encouraging a common embedding geometry. Because image and text frontends produce features with different statistical structures, however, complete parameter sharing may be overly restrictive—motivating a shared quantum backbone with lightweight modality-specific adaptation.

Robust retrieval across variational training runs. Different initializations may produce embedding spaces with similar aggregate quality but different local neighborhoods, leading to changes in the ordering of retrieved candidates. Although this variability reduces the reliability of any single model, it can also produce complementary rankings across independently trained models.

Ensembles improve robustness by combining models with complementary errors, with effectiveness depending on both member strength and diversity Dietterich (2000); Krogh & Vedelsby (1995); Brown et al. (2005). In quantum machine learning, ensembles have been explored primarily for classification Schuld & Petruccione (2018); Li et al. (2024); Tolotti et al. (2024); West et al. (2023), combining class predictions rather than aggregating ranked candidates.

Together, the preceding work motivates the two components of QMORE. To use a constrained variational parameter budget efficiently, QMORE shares one quantum encoder across image and text inputs while retaining lightweight modality-specific phase adapters. To reduce sensitivity to individual training runs, it combines independently trained models using an accuracy–diversity weighted rank-fusion rule. The next section formalizes the retrieval pipeline and introduces each component.

3 METHOD

We first formalize a contrastive baseline for quantum image–text retrieval, denoted *Quantum Baseline*, which adapts recent quantum multimodal frameworks Chen et al. (2025); Zheng et al. (2025) to paired image–text retrieval.

3.1 RETRIEVAL PIPELINE AND TRAINING OBJECTIVE

Figure 1 (left) illustrates the quantum baseline. For each paired sample $(x_i^{\text{img}}, x_i^{\text{text}})$, the pipeline produces a quantum-ready feature vector per modality using classical preprocessing and a lightweight projection head. Images are processed with pretrained weights (e.g., ResNet), and each modality m is mapped by a lightweight projection head f_{ϕ_m} to a classical embedding $\mathbf{h}_i^m = f_{\phi_m}(x_i^m) \in \mathbb{R}^d$ for $m \in \{\text{img}, \text{text}\}$, where d matches the input dimension required by the quantum embedding layer. Each modality is then encoded into a quantum state by its own variational quantum encoder.

The quantum baseline uses *separate* variational quantum encoders per modality, $U_{\theta_{\text{img}}}$ and $U_{\theta_{\text{text}}}$. The modality feature $\mathbf{h}_i^m$ is encoded with angle embedding $E(\cdot)$ into an n-qubit state, followed by L variational layers of parameterized single-qubit rotations (e.g., R_x) and CNOTs for entanglement. The state is measured to produce an embedding $\mathbf{z}_i^m \in \mathbb{R}^{d_z}$ used for retrieval. The two modality embeddings are trained with a contrastive objective that pulls matched image–text pairs together and pushes mismatched pairs apart.

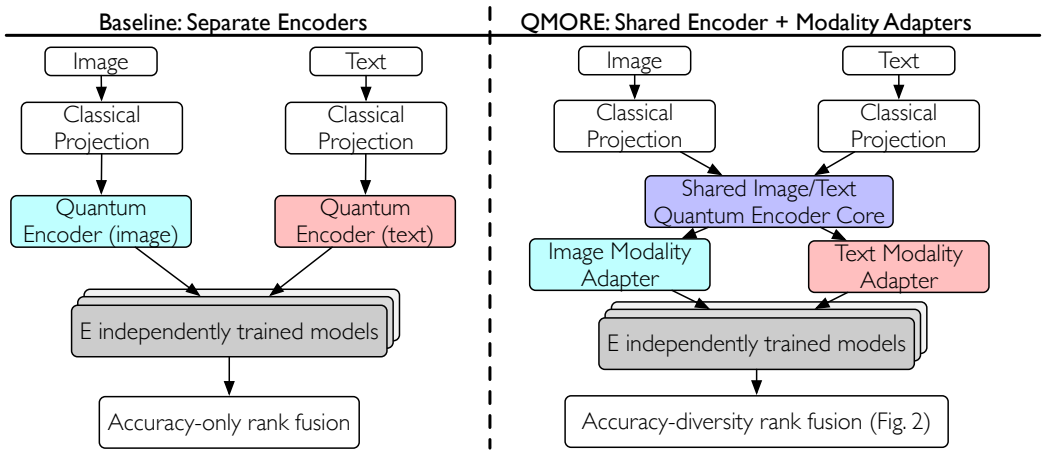


Figure 1: **Baseline vs. QMORE.** Left: the quantum baseline assigns a separate variational encoder to each modality, duplicating circuit parameters. Right: QMORE concentrates capacity in a single shared quantum encoder, with per-modality flexibility provided by n pre-measurement phase rotations per modality (Eq. 2). Both pipelines ensemble E independently trained models; baselines use accuracy-only rank fusion, while QMORE uses its accuracy–diversity rank fusion.

Contrastive training objective. The quantum baseline is trained end-to-end with the InfoNCE loss Oord et al. (2018); Radford et al. (2021): within each batch, the matched caption is the positive for its image and all other in-batch captions are negatives, under cosine similarity with temperature τ. At inference, each ensemble member ranks candidates by cosine similarity. These rankings are combined using rank fusion, and the final ranking is evaluated with Recall@K and rank statistics.

While this baseline provides a natural quantum extension of contrastive retrieval, it inherits the two NISQ-era limitations identified above: parameter duplication across modalities and run-to-run training variability. QMORE retains the same contrastive pipeline and addresses each with the two techniques shown in Figure 1 (right): a shared quantum encoder with lightweight modality adapters, and an accuracy–diversity weighted rank ensemble, described next.

3.2 Shared Quantum Encoder with Modality Adapters

In contrast to the quantum baseline's separate encoders, QMORE reuses a single quantum encoder across modalities, concentrating capacity into one shared embedding geometry with modality-specific flexibility handled by lightweight adapters.

QMORE's Shared encoder. Let f_{ϕ_m} denote the classical preprocessing and projection head for modality $m \in \{\text{img}, \text{text}\}$, producing $\mathbf{h}_i^m \in \mathbb{R}^d$ as above. Both modalities are then embedded by the *same* variational circuit U_θ:

$$|\psi_i^m\rangle \;=\; U_\theta\, E(\mathbf{h}_i^m)\,|0^n\rangle\,, \qquad m \in \{\text{img}, \text{text}\}, \tag{1}$$

where $E(\cdot)$ denotes the fixed data-encoding map and θ is shared across modalities. Sharing U_θ makes both modalities obey the same embedding map, so contrastive learning directly shapes a single shared geometry rather than reconciling two independently learned feature maps. In classical multimodal and transfer settings, parameter-efficient adaptation often adds small trainable modules to a shared backbone (e.g., adapters or low-rank updates) Houlsby et al. (2019); Hu et al. (2021). QMORE extends this principle to the quantum encoder by introducing lightweight *modality adapters* as small pre-measurement phase rotations that adjust the effective measurement basis for each modality, providing modality-specific flexibility with minimal additional parameters.

Modality adapters. QMORE applies the modality adapter A_{α_m} before measurement:

$$U_\theta^m \;=\; A_{\alpha_m}\, U_\theta, \qquad A_{\alpha_m} \;=\; \prod_{k=1}^{n} R_Z^{(k)}(\alpha_{m,k}), \tag{2}$$

where n is the number of qubits, $R_Z^{(k)}(\cdot)$ is a Z rotation on qubit k, and $\alpha_m \in \mathbb{R}^n$ are trainable per-qubit phases. Because embeddings are read out using Pauli-Y expectations, these phase rotations change the effective measurement axes and therefore provide modality-specific adaptation despite being applied immediately before measurement. This adds only n parameters per modality. The adapters apply modality-specific phase offsets before measurement, providing flexibility while keeping the shared encoder U_θ the same across modalities. Compared to two separate encoders $(U_{\theta_{\text{img}}}, U_{\theta_{\text{text}}})$, QMORE reuses a single quantum encoder U_θ for both modalities and learns only n additional adapter angles per modality. This is appealing in resource-limited NISQ regimes, where variational optimization can be fragile McClean et al. (2018); Larocca et al. (2025), and aligns with evidence that parameter sharing can preserve alignment while improving efficiency You et al. (2022). We next address run-to-run variability.

3.3 Accuracy–Diversity Rank Fusion

Independently trained variational models can converge to different solutions, producing different embedding neighborhoods and retrieval orderings across runs McClean et al. (2018); Cerezo et al. (2021); Grant et al. (2019). QMORE therefore ensembles independently trained embedding models so that complementary solutions reinforce one another Dietterich (2000); Krogh & Vedelsby (1995).

Designing an effective ensemble for retrieval is nontrivial: naively averaging similarity scores can be brittle as different models may output scores on different scales, even when their within-model rankings are reasonable. Conversely, selecting only the best validation checkpoint collapses the ensemble to a single member and discards complementary solutions. To address this, QMORE uses a retrieval-aware ensemble (Fig. 2 in Appendix A) that (i) weights members by retrieval strength, (ii) promotes diversity by favoring members whose embeddings add complementary information, and (iii) aggregates predictions with rank-based fusion to optimize retrieval ordering.

Rank-based ensemble aggregation. For a retrieval query q (image→text in our evaluation), each ensemble member $e \in \{1, \ldots, E\}$ produces a ranked list over the C candidates. Let $r_e(q, c) \in \{0, 1, \ldots, C-1\}$ denote the rank position assigned to candidate c by member e, with $r_e(q, c) = 0$ indicating the top prediction. Because independently trained quantum models can produce similarity scores on different scales, QMORE aggregates predictions using rank information, which depends only on the induced ordering. We convert ranks into a vote score $s_e(q, c)$ via an exponential decay, $s_e(q, c) = \exp\big(-\lambda\, r_e(q, c)\big)$, where $\lambda > 0$ controls how sharply the ensemble emphasizes top-ranked candidates.

Weighting ensemble members by retrieval strength and diversity. Rank-based votes determine how each member scores candidates, yet equal weighting is rarely optimal: some runs produce stronger retrieval orderings, while others are redundant. QMORE therefore assigns each member e a scalar *retrieval-strength weight* π_e (the "accuracy" in accuracy–diversity), computed on a fixed calibration subset, and a scalar *diversity score* δ_e that reflects how non-redundant its embeddings are relative to other members. These signals are fused into an ensemble weight w_e, which controls the influence of member e in the aggregated ranking.

Retrieval-strength weight. For each member e, we compute the mean rank-decayed true-match score on the N calibration pairs, $\tilde{\pi}_e = N^{-1} \sum_{i=1}^{N} \exp(-\lambda r_e(i, i))$. We convert these scores into relative weights via

$$\pi_e = \mathrm{softmax}(\mathrm{minmax}(\tilde{\pi}_e)/\tau_\pi)\,,$$

where normalization is across members and $\tau_\pi = 0.07$.

Diversity score. High retrieval strength alone is not sufficient: members can achieve similar calibration performance yet induce nearly identical cross-modal distance structures, limiting ensemble gains Kuncheva & Whitaker (2003); Brown et al. (2005). To favor complementary members, QMORE measures how differently each member organizes the calibration pairs.

For the N calibration pairs, member e induces the cross-modal cosine-distance matrix

$$\mathbf{D}_e[a, b] = 1 - \cos\big(\mathbf{z}_{a,e}^{\text{img}}, \mathbf{z}_{b,e}^{\text{text}}\big)\,,$$

where rows correspond to image queries and columns to text candidates. We summarize this structure by

$$\mathbf{v}_e = \mathrm{vec}(\{\mathbf{D}_e[a, b] : 1 \le a < b \le N\})\,,$$

using the same fixed subset of off-diagonal entries for every member. For members (e, e'), their distance-structure dissimilarity is

$$d(e, e') = 1 - \mathrm{corr}(\mathbf{v}_e, \mathbf{v}_{e'}),$$

where $\mathrm{corr}(\cdot, \cdot)$ denotes Pearson correlation. This comparison measures disagreement between members' cross-modal distance structures rather than differences between their embedding coordinates. The diversity score for member e is its mean dissimilarity to the remaining ensemble members, normalized and rescaled:

$$\delta_e = \frac{1}{2} + \frac{1}{2}\,\mathrm{minmax}\left(\frac{1}{E-1}\sum_{e' \neq e} d(e, e')\right), \tag{3}$$

where $\mathrm{minmax}(\cdot)$ denotes min–max normalization across the E ensemble members. This affine rescaling maps the normalized diversity score from $[0, 1]$ to $[0.5, 1]$, preventing low-diversity members from receiving near-zero diversity weight. Larger δ_e indicates a less redundant cross-modal distance structure and therefore greater potential utility to the ensemble.

Final weighted rank aggregation. We combine retrieval strength and diversity using $w_e = \rho\,\pi_e + (1 - \rho)\,\delta_e$, where $\rho \in [0, 1]$ shifts emphasis from diversity toward retrieval strength. A common normalization of $\{w_e\}$ is immaterial to the resulting ranking. We score candidate c for query q by

$$S(q, c) = \sum_{e=1}^{E} w_e \exp\big(-\lambda r_e(q, c)\big). \tag{4}$$

Candidates are returned in descending order of $S(q, c)$.

4 Experimental Setup

Datasets. We evaluate image-to-text retrieval on three paired vision–language datasets spanning controlled to in-the-wild settings: **SVO Probes** Hendricks & Nematzadeh (2021), our **Captioned CIFAR-10 (CCIFAR)**, and **Flickr30k** Young et al. (2014). SVO Probes provides controlled subject–verb–object captions, isolating compositional alignment from visual and linguistic complexity; CCIFAR pairs CIFAR-10 images Krizhevsky (2009) with GPT-4o-mini-generated captions, giving a low-resolution benchmark with a controlled caption distribution; Flickr30k tests transfer to natural images with free-form human captions. SVO Probes and CCIFAR each contain 8,000 training and 2,000 test pairs (240×240 and 32×32 images, respectively); Flickr30k uses 28,000 training images (140,000 image–caption pairs) and 2,000 held-out pairs, with images resized and center-cropped to 224×224.

Shared pipeline. Images are encoded with a frozen ResNet-18 (512-d penultimate features) and captions with a frozen BERT-tiny encoder (`prajjwal1/bert-tiny`; 128-d); features are standardized with training-split statistics. Every model then applies a trainable linear projection followed by a tanh activation to produce an 8-d encoder input; this compressor is identical in form across all models and excluded from the parameter-matching budget below. Quantum circuits are simulated classically in PennyLane Bergholm et al. (2022); classical components use PyTorch.

Quantum models. QMORE uses n=8 qubits and 8 layers of PennyLane `BasicEntanglerLayers`, reading out per-qubit Pauli-Y expectations as the embedding. The quantum baseline uses the same circuit family and readout but a dual-encoder design; each of its two encoders uses 5 layers so that their combined parameter count matches QMORE's replacement-block budget. Throughout, parameter matching counts only the *replacement block*: QMORE's 64 circuit parameters plus 16 adapter phases (80 total).

Classical baselines. Two parameter-matched classical variants replace the quantum encoder core with a classical alignment head while retaining the same trainable input projection. The *separate* baseline mirrors the quantum baseline's dual-encoder design, splitting the replacement budget across independent image and text MLPs; the *shared* baseline mirrors QMORE, sharing one linear encoder across modalities with additive modality-specific bias adapters, isolating whether QMORE's gains come from the sharing principle alone or from the quantum encoder itself. The shared baseline is

also the strongest classical configuration we found at this budget (Appendix C.3). For the parameter-scaling study, we use classical replacement blocks at approximately $1\times$ and $5\times$ QMORE's 80-parameter budget (78–80 and 398–400 parameters, respectively), holding the loss, input features, optimization, and training budget fixed.

Training. All methods are trained with InfoNCE (temperature 0.07) using Adam at learning rate 0.003 for 1,000 steps, each on a randomly sampled minibatch of 64 pairs.

Ensembling and fusion. Unless otherwise stated, results use 10-member ensembles of identical architectures, with members obtained from independent stochastic training runs that differ in parameter initialization and minibatch sampling; prior quantum multimodal pipelines use single models, so we equip every baseline with an equally sized ensemble for fair comparison. QMORE uses its novel accuracy–diversity rank fusion; quantum and classical baselines use accuracy-only rank fusion. All calibration quantities are computed on a 500-pair subset of the training pool, disjoint from all evaluation data: member weights are estimated there, and QMORE's ρ=0.3 (favoring diversity) and λ=0.05 (rank decay) were selected once on this split and applied unchanged to every dataset.

Noisy simulation. To assess robustness to device noise, we re-evaluate the Flickr30k models under a calibrated noise model of IBM's 16-qubit `ibmq_guadalupe`, following common practice for parameterized circuits Wang et al. (2022a;b); backend details are given in Appendix D. Each 8-qubit circuit is executed with 2,048 shots; models are trained in ideal simulation and evaluated under noise without retraining, with all other pipeline components unchanged.

Evaluation protocol. All methods use identical splits and protocol. The 2,000 test pairs are divided into 10 fixed chunks of 200 queries, drawn with a shared seed and identical across models to enable paired per-chunk comparisons; for datasets with multiple captions per image, each chunk contains at most one caption per image. Each ensemble member ranks candidates by cosine similarity. The final ranking is obtained using the corresponding rank-fusion rule. We report Recall@K and the mean and median rank of the ground-truth caption as mean $\pm$ std over chunks.

5 Results and Analysis

5.1 Retrieval Performance and Parameter Scaling

We compare QMORE with the quantum dual-encoder and two parameter-matched classical baselines under identical training, feature, and evaluation protocols. On SVO Probes, QMORE reaches **15.8% R@1** versus **9.0%** (quantum) and **12.6%** (best classical), a **+3.2** point gain over the strongest baseline at the strictest recall level. Additionally, these gains extend from controlled probes to real-world captioned data. On Flickr30k, QMORE improves **R@1 to 14.4%** (vs. **3.4%** quantum, **11.6%** shared classical), indicating that the shared-geometry inductive bias transfers beyond synthetic and small datasets. Table 1 summarizes results across datasets: QMORE improves *every* metric we report over all baselines, suggesting that in resource-constrained regimes, *how* quantum parameters are used matters. Notably, the shared classical baseline substantially outperforms its separate counterpart (e.g., **12.6%** vs. **5.0%** R@1 on SVO Probes), confirming the value of the sharing principle itself, yet QMORE improves further on every metric, indicating that the quantum encoder contributes beyond parameter sharing alone.

This trend persists when the classical encoders are scaled from approximately **1**$\times$ to **5**$\times$ QMORE's parameter budget (Tab. 2). Across all three datasets, QMORE remains best on every reported metric. On Flickr30k, for example, it reaches **40.9% R@5**, compared with **36.8%** for the **5**$\times$ shared classical encoder. The same pattern holds for the separate architecture, suggesting that the quantum encoder provides a useful inductive bias for compact retrieval embeddings.

5.2 Component Ablations

QMORE uses a shared quantum backbone with modality adapters and learns a single image–text embedding geometry. When we remove the shared encoder and use modality-specific parameterization, retrieval drops sharply. On SVO Probes, R@1 decreases from **15.8%** to **10.1%** (Tab. 3), showing that the gains come from shared structure rather than added capacity. The trend holds on Flickr30k: the full model reaches **14.4% R@1** versus **6.8%** without the shared encoder, and re-

Table 1: Retrieval performance of QMORE vs. parameter-matched baselines (mean ± std over 10 chunks; Recall@K in %). QMORE uses its novel accuracy–diversity fusion method; baselines use accuracy-only rank fusion. Shared classical models include modality-specific adapters. ↑/↓: higher/lower better.

Dataset	Model	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓	Median ↓
SVO Probes	Classical (separate)	5.0±1.4	17.5±2.2	29.3±2.8	32.5±2.1	19.7±2.2
	Quantum (separate)	9.0±1.7	27.5±1.7	40.8±2.7	25.4±1.6	13.4±1.7
	Classical (shared)	12.6±1.6	38.7±3.3	55.4±3.0	17.7±1.7	7.5±1.0
	QMORE	**15.8±2.0**	**44.6±2.7**	**61.6±1.9**	**16.1±1.5**	**5.6±0.5**
CCIFAR	Classical (separate)	4.3±1.3	19.2±2.2	37.8±2.6	22.3±1.8	13.6±0.9
	Quantum (separate)	6.0±1.1	24.5±2.9	44.7±1.7	18.2±1.4	11.1±0.6
	Classical (shared)	8.5±1.8	35.7±2.1	56.0±2.6	13.9±1.2	8.0±0.6
	QMORE	**10.7±1.9**	**39.4±3.1**	**60.1±2.5**	**12.5±1.1**	**6.7±0.7**
Flickr 30k	Classical (separate)	3.5±1.1	16.0±2.9	28.2±2.1	38.5±1.5	26.5±2.5
	Quantum (separate)	3.4±1.2	15.7±2.1	27.4±2.3	38.6±2.4	25.3±2.6
	Classical (shared)	11.6±1.2	35.5±3.0	50.5±2.4	21.0±1.9	9.2±1.3
	QMORE	**14.4±1.2**	**40.9±2.0**	**56.1±3.0**	**19.9±1.9**	**7.3±0.6**

Table 2: Encoder-core scaling: classical baselines use replacement blocks at 1× and 5× QMORE's 80-parameter encoder-core budget (78–80 and 398–400 parameters, respectively); all other components are held fixed (see Section 4). Results are reported as mean ± standard deviation over 10 chunks; Recall@K is reported in %, and mean and median are reported as rank positions. QMORE performs best throughout. ↑/↓: higher/lower is better.

Dataset	Model	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓	Median ↓
SVO Probes	Classical (separate, 1×)	5.0±1.4	17.5±2.2	29.3±2.8	32.5±2.1	19.7±2.2
	Classical (separate, 5×)	12.3±1.6	40.2±1.6	57.8±2.3	17.6±2.2	7.1±0.5
	Classical (shared, 1×)	12.6±1.6	38.7±3.3	55.4±3.0	17.7±1.7	7.5±1.0
	Classical (shared, 5×)	10.5±1.9	38.6±2.3	57.1±2.9	17.4±1.8	6.9±0.9
	QMORE	**15.8±2.0**	**44.6±2.7**	**61.6±1.9**	**16.1±1.5**	**5.6±0.5**
CCIFAR	Classical (separate, 1×)	4.3±1.3	19.2±2.2	37.8±2.6	22.3±1.8	13.6±0.9
	Classical (separate, 5×)	7.6±1.7	31.2±2.2	52.8±1.7	15.5±1.2	8.8±0.6
	Classical (shared, 1×)	8.5±1.8	35.7±2.1	56.0±2.6	13.9±1.2	8.0±0.6
	Classical (shared, 5×)	9.3±2.1	36.3±3.6	57.4±3.7	12.9±1.3	7.5±0.9
	QMORE	**10.7±1.9**	**39.4±3.1**	**60.1±2.5**	**12.5±1.1**	**6.7±0.7**
Flickr 30k	Classical (separate, 1×)	3.5±1.1	16.0±2.9	28.2±2.1	38.5±1.5	26.5±2.5
	Classical (separate, 5×)	10.7±1.6	34.4±3.0	48.7±2.2	21.0±1.7	10.2±1.0
	Classical (shared, 1×)	11.6±1.2	35.5±3.0	50.5±2.4	21.0±1.9	9.2±1.3
	Classical (shared, 5×)	12.8±1.8	36.8±3.5	50.2±2.7	21.6±1.9	9.4±1.1
	QMORE	**14.4±1.2**	**40.9±2.0**	**56.1±3.0**	**19.9±1.9**	**7.3±0.6**

moving the ensemble reduces R@1 to **8.7%**. On SVO Probes and Flickr30k, the largest relative improvements are at **R@1**, consistent with better top-rank alignment from shared parameterization and rank fusion.

5.3 Ensemble-Method Comparison

We compare retrieval ensemble rules (top-1 majority voting and accuracy-only rank fusion) to QMORE's accuracy–diversity rank fusion (Tab. 4). The method achieves the best Recall@K on Flickr30k: **14.4/40.9/56.1** vs. 13.3/35.4/52.7 for top-1 voting. Notably, even under the baselines' accuracy-only rank fusion rule, QMORE outperforms all Flickr30k baselines in Table 1 on Recall@K (12.8% R@1 vs. 11.6% for the best baseline), so its advantage is not only an artifact of the stronger fusion.

Table 3: QMORE ablations (mean ± std over 10 chunks). Recall@K in %; mean and median are rank positions. "w/o shared encoder" uses modality-specific parameters; "w/o ensemble" reports mean single-member performance. ↑/↓: higher/lower better.

Dataset	Variant	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓	Median ↓
SVO Probes	w/o shared encoder	10.1±1.1	33.6±2.9	47.7±2.7	21.9±1.5	10.6±1.3
	w/o ensemble	8.5±0.6	30.2±1.2	46.6±1.7	23.0±1.5	10.9±0.7
	QMORE	**15.8±2.0**	**44.6±2.7**	**61.6±1.9**	**16.1±1.5**	**5.6±0.5**
CCIFAR	w/o shared encoder	9.0±1.2	30.3±2.0	48.9±2.8	15.6±1.1	9.8±0.6
	w/o ensemble	6.8±0.6	29.0±1.4	49.0±2.0	17.1±1.1	9.8±0.5
	QMORE	**10.7±1.9**	**39.4±3.1**	**60.1±2.5**	**12.5±1.1**	**6.7±0.7**
Flickr 30k	w/o shared encoder	6.8±1.8	23.1±2.0	36.6±2.0	35.6±2.3	17.8±1.4
	w/o ensemble	8.7±0.7	29.1±1.1	43.8±1.5	26.4±1.7	12.6±0.9
	QMORE	**14.4±1.2**	**40.9±2.0**	**56.1±3.0**	**19.9±1.9**	**7.3±0.6**

Accuracy–diversity fusion also improves both classical baselines on Flickr30k (Tab. 4), showing that the ensemble rule is not quantum-specific. With accuracy–diversity fusion applied to both models, QMORE achieves 14.4% versus 13.5% R@1 for the strongest classical configuration, with significant differences under a paired t-test (p = 0.030) and Wilcoxon signed-rank test (p = 0.042).

Table 4: Ensemble-method comparison on Flickr30k for QMORE and the parameter-matched classical baselines (mean ± std over 10 chunks; best per block in bold). "Acc. ens." is the fusion used by baselines in Tab. 1. ↑/↓: higher/lower is better.

Model	Method	R@1 ↑	R@5 ↑	R@10 ↑	Mean ↓	Median ↓
Classical (separate)	Top-1 maj.	3.9±1.3	14.7±1.8	24.7±2.1	**37.8±2.1**	23.8±2.3
	Acc. ens.	3.5±1.1	16.0±2.9	28.2±2.1	38.5±1.5	26.5±2.5
	Acc-Div ens.	**4.6±0.9**	**17.5±2.0**	**29.2±2.0**	38.9±1.7	**23.7±1.9**
Classical (shared)	Top-1 maj.	13.0±1.4	35.9±2.4	53.5±1.7	19.7±2.0	8.2±0.6
	Acc. ens.	11.6±1.2	35.5±3.0	50.5±2.4	21.0±1.9	9.2±1.3
	Acc-Div ens.	**13.5±1.6**	**39.5±4.0**	**54.7±2.4**	**19.5±2.1**	**7.6±0.8**
QMORE	Top-1 maj.	13.3±1.6	35.4±2.9	52.7±2.7	20.2±2.0	8.8±1.0
	Acc. ens.	12.8±1.9	38.3±3.8	53.2±3.1	21.5±1.9	8.4±1.2
	Acc-Div ens.	**14.4±1.2**	**40.9±2.0**	**56.1±3.0**	**19.9±1.9**	**7.3±0.6**

5.4 Robustness to Simulated Noise

Under the calibrated `ibmq_guadalupe` noise model (see Section 4), QMORE degrades from 14.4/40.9/56.1 to 13.3/38.9/54.7 R@1/R@5/R@10, retaining 92% of its R@1. The quantum baseline instead drops from 3.4/15.7/27.4 to 0.4/3.2/5.6 (near random), showing that QMORE's advantage persists under realistic device noise.

6 Conclusion

We introduced QMORE, a hybrid quantum–classical framework for image–text retrieval that enforces a shared cross-modal geometry through a single variational quantum encoder with lightweight modality adapters, and improves robustness by fusing independent runs with an accuracy–diversity weighted rank ensemble. Across three benchmarks, QMORE outperforms parameter-matched quantum and classical baselines and retains its advantage under a calibrated noise model of a near-term IBM device. We acknowledge that our retrieval performance remains well below that of large-scale systems such as CLIP—a constraint of the low-parameter regime, not a claim to compete at scale. We instead show that *how* quantum parameters are allocated matters: a shared multimodal geometry with diversity-aware ensembling extracts more retrieval signal from a fixed quantum budget than modality-specific or unensembled alternatives.

AI USE STATEMENT

We used ChatGPT to suggest revisions and phrasing, assist in drafting small portions of manuscript sections, suggest relevant references, and assist with code review and debugging. We used GPT-4o-mini to generate captions for existing CIFAR-10 images in CCIFAR; this was our only use of generative AI to generate synthetic data. The caption-generation, normalization, and filtering procedures are described in Appendix B. We checked the suggested references and reviewed and tested the AI-assisted code. The authors take full responsibility for the final manuscript and associated research artifacts.

REPRODUCIBILITY STATEMENT

We provide implementation code and instructions for training and evaluating QMORE and its classical baselines in the supplementary artifact. Appendix C documents the architectures, parameter counts, feature preprocessing, training hyperparameters, ensemble and evaluation protocols, random-seed handling, and software and hardware environment. Appendix B describes the construction of Captioned CIFAR-10, including caption generation, normalization, and filtering. Appendix D specifies the backend, noise model, and simulation settings used in the robustness study. The artifact README describes the required external data, feature formats, dependencies, and commands for training and evaluation.

REFERENCES

Ville Bergholm, Josh Izaac, Maria Schuld, Christian Gogolin, Shahnawaz Ahmed, Vishnu Ajith, M. Sohaib Alam, Guillermo Alonso-Linaje, B. AkashNarayanan, Ali Asadi, Juan Miguel Arrazola, Utkarsh Azad, Sam Banning, Carsten Blank, Thomas R Bromley, Benjamin A. Cordier, Jack Ceroni, Alain Delgado, Olivia Di Matteo, Amintor Dusko, Tanya Garg, Diego Guala, Anthony Hayes, Ryan Hill, Aroosa Ijaz, Theodor Isacsson, David Ittah, Soran Jahangiri, Prateek Jain, Edward Jiang, Ankit Khandelwal, Korbinian Kottmann, Robert A. Lang, Christina Lee, Thomas Loke, Angus Lowe, Keri McKiernan, Johannes Jakob Meyer, J. A. Montañez-Barrera, Romain Moyard, Zeyue Niu, Lee James O'Riordan, Steven Oud, Ashish Panigrahi, Chae-Yeun Park, Daniel Polatajko, Nicolás Quesada, Chase Roberts, Nahum Sá, Isidor Schoch, Borun Shi, Shuli Shu, Sukin Sim, Arshpreet Singh, Ingrid Strandberg, Jay Soni, Antal Száva, Slimane Thabet, Rodrigo A. Vargas-Hernández, Trevor Vincent, Nicola Vitucci, Maurice Weber, David Wierichs, Roeland Wiersema, Moritz Willmann, Vincent Wong, Shaoming Zhang, and Nathan Killoran. Pennylane: Automatic differentiation of hybrid quantum-classical computations, 2022. URL https://arxiv.org/abs/1811.04968.

Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20, 2005.

M. Cerezo, A. Arrasmith, R. Babbush, S. C. Benjamin, S. Endo, K. Fujii, J. R. McClean, K. Mitarai, X. Yuan, L. Cincio, and P. J. Coles. Variational quantum algorithms. *Nature Reviews Physics*, 2021. doi: 10.1038/s42254-021-00348-9.

Chi-Sheng Chen, Aidan Hung-Wen Tsai, and Sheng-Chieh Huang. Quantum multimodal contrastive learning framework. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889504.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems (MCS 2000)*, volume 1857 of *Lecture Notes in Computer Science*, pp. 1–15. Springer, 2000. doi: 10.1007/3-540-45014-9_1.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. URL https://bmva-archive.org.uk/bmvc/2018/contents/papers/0344.pdf.

Edward Grant, Leonard Wossnig, Mateusz Ostaszewski, and Marcello Benedetti. An initialization strategy for addressing barren plateaus in parametrized quantum circuits. *Quantum*, 3:214, 2019. doi: 10.22331/q-2019-12-09-214.

Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019. doi: 10.1038/s41586-019-0980-2. URL https://arxiv.org/abs/1804.11326.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3635–3644, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.318. URL https://aclanthology.org/2021.findings-acl.318/.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. URL https://www.jair.org/index.php/jair/article/view/10833.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning (ICML)*, 2019. URL https://arxiv.org/abs/1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. URL https://arxiv.org/abs/2106.09685.

Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean. Power of data in quantum machine learning. *Nature Communications*, 12(1), May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22539-9. URL http://dx.doi.org/10.1038/s41467-021-22539-9.

Ali Javadi-Abhari, Matthew Treinish, Kevin Krsulich, Christopher J. Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D. Nation, Lev S. Bishop, Andrew W. Cross, Blake R. Johnson, and Jay M. Gambetta. Quantum computing with Qiskit, 2024.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 4904–4916. PMLR, 2021. URL https://arxiv.org/abs/2102.05918.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137, 2015. URL https://openaccess.thecvf.com/content_cvpr_2015/papers/Karpathy_Deep_Visual-Semantic_Alignments_2015_CVPR_paper.pdf.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. URL https://proceedings.mlr.press/v202/koh23a/koh23a.pdf.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7, pp. 231–238, 1995. URL https://proceedings.neurips.cc/paper/1994/hash/b8c37e33defde51cf91e1e03e51657da-Abstract.html.

Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

Martin Larocca, Supanut Thanasilp, Samson Wang, Kunal Sharma, Jacob Biamonte, Patrick J Coles, Lukasz Cincio, Jarrod R McClean, Zoë Holmes, and Marco Cerezo. Barren plateaus in variational quantum computing. *Nature Reviews Physics*, 7(4):174–189, 2025.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *European Conference on Computer Vision (ECCV)*, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Kuang-Huei_Lee_Stacked_Cross_Attention_ECCV_2018_paper.html.

Qingyu Li, Yuhan Huang, Xiaokai Hou, Ying Li, Xiaoting Wang, and Abolfazl Bayat. Ensemble-learning error mitigation for variational quantum shallow-circuit classifiers. *Physical Review Research*, 6(1):013027, 2024. doi: 10.1103/PhysRevResearch.6.013027.

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), November 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-07090-4. URL http://dx.doi.org/10.1038/s41467-018-07090-4.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. URL https://openaccess.thecvf.com/content_iccv_2015/html/Plummer_Flickr30k_Entities_Collecting_ICCV_2015_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://arxiv.org/abs/2103.00020.

Maria Schuld and Nathan Killoran. Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019. doi: 10.1103/PhysRevLett.122.040504. URL https://arxiv.org/abs/1803.07128.

Maria Schuld and Francesco Petruccione. Quantum ensembles of quantum classifiers. *Scientific Reports*, 8(1):2772, 2018. doi: 10.1038/s41598-018-20403-3.

Emiliano Tolotti, Enrico Zardini, Enrico Blanzieri, and Davide Pastorello. Ensembles of quantum classifiers. *Quantum Information and Computation*, 24(3–4):181–209, 2024. doi: 10.26421/QIC24.3-4-1.

Hanrui Wang, Yongshan Ding, Jiaqi Gu, Yujun Lin, David Z. Pan, Frederic T. Chong, and Song Han. Quantumnas: Noise-adaptive search for robust quantum circuits. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 692–708. IEEE, April 2022a. doi: 10.1109/hpca53966.2022.00057. URL http://dx.doi.org/10.1109/HPCA53966.2022.00057.

Hanrui Wang, Jiaqi Gu, Yongshan Ding, Zirui Li, Frederic T Chong, David Z Pan, and Song Han. Quantumnat: quantum noise-aware training with noise injection, quantization and normalization. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, pp. 1–6, 2022b.

Maxwell T. West, Martin Sevior, and Muhammad Usman. Boosted ensembles of qubit and continuous variable quantum support vector machines for B meson flavor tagging. *Advanced Quantum Technologies*, 6(10):2300130, 2023. doi: 10.1002/qute.202300130.

Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. URL https://arxiv.org/abs/2207.12661.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Jin Zheng, Qing Gao, Daoyi Dong, Jinhu Lü, and Yue Deng. A quantum multimodal neural network model for sentiment analysis on quantum circuits. *IEEE Transactions on Artificial Intelligence*, 6 (5):1128–1142, 2025. doi: 10.1109/TAI.2024.3511514.

Wangchunshu Zhou, Yan Zeng, Shizhe Diao, and Xinsong Zhang. VLUE: A multi-task multi-dimension benchmark for evaluating vision-language pre-training. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhou22n.html.

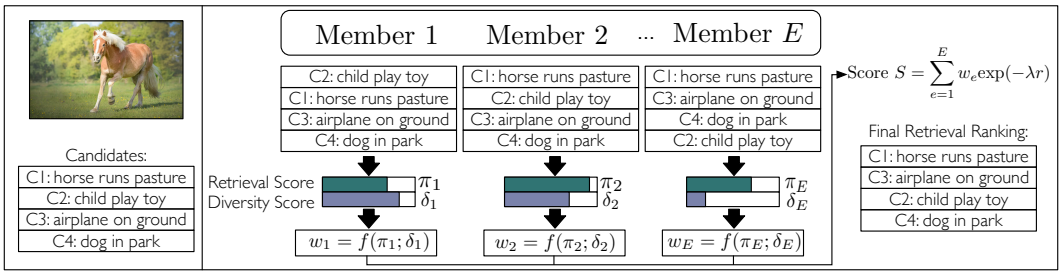


Figure 2: QMORE's accuracy–diversity rank fusion ensemble. Members are weighted by retrieval strength π_e and embedding diversity δ_e via $w_e = \rho\,\pi_e + (1-\rho)\,\delta_e$, then combined by rank-decayed voting (Eq. 4).



Figure 3: Example image–caption pairs from Captioned CIFAR-10 (CCIFAR). Each image is paired with a three-word caption generated by `gpt-4o-mini`. We also show the underlying CIFAR-10 class label for reference.

A Overview

This appendix provides the implementation details needed to reproduce the QMORE experimental setup and all baseline configurations. It covers: (i) the construction of our Captioned CIFAR-10 (CCIFAR) benchmark, including caption generation, normalization, and filtering; (ii) the exact architecture and trainable-parameter accounting for every model—the shared feature frontend, QMORE's quantum encoder with a full derivation of the 80-parameter budget, the separate quantum baseline, and the classical parameter-matched baselines at $1\times$ and $5\times$ scale—together with the training, ensembling, evaluation, random-seed handling, feature standardization, and software/hardware environment; and (iii) reproducibility details for the noisy-simulation study, including the `FakeGuadalupeV2` backend and its static calibration snapshot, native gate set, connectivity, noise channels, shot count, and absence of error mitigation. We emphasize that all "noisy" results are produced by classical simulation under an IBM hardware-derived noise model, not by execution on physical quantum hardware.

B Captioned CIFAR-10 Construction Details

To construct Captioned CIFAR-10 (CCIFAR), we augment CIFAR-10 images with generated captions. Our goal is to create a compact paired image–text retrieval benchmark with simple captions that remain lightweight enough for controlled low-parameter experiments.

B.1 Dataset Size and Splits

CCIFAR is constructed directly from the standard CIFAR-10 training and test partitions. In our experiments, we use 8,000 image–caption pairs for training and 2,000 for testing.

B.2 Caption Generation Procedure

We generate one caption per image using a vision-capable language model. For each CIFAR-10 image, we prompt the model with the image and a strict output constraint:

> *Exactly 3 lowercase words. No punctuation. No articles. No numbers. No "photo" "image" "picture". Output only the 3 words.*

This format was chosen to keep captions short, uniform, and easy to parse, while still providing simple semantic descriptions aligned with the image content. We use `gpt-4o-mini` in all caption-generation experiments. Caption generation is performed asynchronously with concurrent API calls to improve throughput.

B.3 Caption Normalization and Filtering

After generation, captions are normalized to enforce a strict three-word format. Specifically, we:

1. convert all text to lowercase,
2. remove surrounding quotation marks or code formatting,
3. replace non-letter characters with spaces,
4. collapse repeated whitespace,
5. remove banned tokens such as `photo`, `image`, `picture`, `a`, `an`, and `the`.

A caption is accepted only if the final output contains exactly three alphabetic words separated by single spaces.

If the initial model output does not satisfy this constraint, we issue one repair request using a shorter corrective prompt:

> *Fix: exactly 3 lowercase words, no punctuation, output only the 3 words.*

If the repaired output still fails validation, we replace it with a deterministic fallback caption template conditioned on the CIFAR-10 class label, such as `small dog scene` or `bright truck object`. This guarantees that every image in CCIFAR is paired with a valid three-word caption.

C Exact Architecture and Parameter Counts

This section specifies every model configuration, including the exact trainable-parameter derivation for the 80-parameter quantum budget and the classical parameter-matched baselines.

C.1 Shared Feature Front-End (all models)

All three datasets—SVO Probes, CCIFAR, and Flickr30k—use the same frozen, pretrained backbones to produce raw modality features:

- **Image:** ResNet-18 (`torchvision`, `ImageNet1K_V1` weights) with the final classification layer replaced by identity; we take the 512**-d** penultimate feature.
- **Text:** BERT-tiny (`prajjwal1/bert-tiny`); we take the 128**-d** `[CLS]` embedding from the last hidden state.

These backbones are *not* trained. Features are cached and then **standardized per feature** with a `StandardScaler` (zero mean, unit variance) whose statistics are *fit on the training split only* and applied unchanged to the test split. Every model—quantum and classical—consumes these identical standardized 512-d / 128-d inputs. Each modality then passes through a trainable Linear$(d \to 8)$ + $\tanh$ compressor before the alignment head.

C.2 QMORE: QUANTUM ENCODER

QMORE maps each standardized feature vector to 8 rotation angles with a per-modality classical *compressor* and then applies a single **shared** 8-qubit variational circuit followed by a per-modality single-qubit adapter and a Pauli-Y readout:

1. **Encoding:** `AngleEmbedding` of the 8 angles as R_X rotations (one per qubit), applied once.
2. **Variational block** (8 layers, basic entangler). Layer $\ell = 1, \ldots, 8$: $R_X(\theta_{\ell,i})$ on every qubit i, followed by a ring of CNOTs.
3. **Modality adapter (RZ):** $R_Z(a_{m,i})$ on every qubit i, with a separate angle vector $a_{\text{img}}, a_{\text{txt}} \in \mathbb{R}^8$ for the two modalities. The circuit weights θ are shared across modalities; *only* the adapter differs.
4. **Readout:** $\langle Y_i \rangle$ on each qubit $\Rightarrow$ an 8-d real embedding. Retrieval uses cosine similarity between image and text embeddings.

The 80-parameter budget (exact derivation). The trainable *quantum* parameters are

$$\underbrace{8 \text{ layers} \times 8 \text{ qubits}}_{R_X \text{ rotation angles}} + \underbrace{8}_{a_{\text{img}}} + \underbrace{8}_{a_{\text{txt}}} = 80.$$

Excluded from the 80: the two compressor layers, $W_{\text{img}}, b_{\text{img}}$ ($512 \times 8 + 8 = 4104$) and $W_{\text{txt}}, b_{\text{txt}}$ ($128 \times 8 + 8 = 1032$), totalling 5136 parameters. These are excluded *not* because they are frozen (they are trained) but because they are a fixed-capacity dimensionality-reduction front-end that is **architecturally identical in every model we compare**. The replacement-block budget therefore isolates the architectural component being compared—the quantum circuit in QMORE versus its classical replacement head—while holding the shared 5136-parameter trainable front-end fixed across the comparison. QMORE's total trainable count is $5136 + 80 = 5216$.

Quantum baseline (separate). The quantum baseline uses the same circuit family and Pauli-Y readout as QMORE but a dual-encoder design: two *independent* 8-qubit circuits (one per modality, no shared weights and no modality adapter), each with 5 layers. Its budget is therefore $2 \times (5 \text{ layers} \times 8 \text{ qubits}) = 80$ circuit parameters—matched to QMORE's 80—atop the same 5136-parameter front-end (total 5216). It isolates the effect of *sharing* the quantum encoder: same quantum capacity, spent on two separate encoders instead of one shared encoder with adapters.

C.3 CLASSICAL PARAMETER-MATCHED BASELINES

Every classical baseline reuses the identical Linear$(d \to 8) + \tanh$ compressor and is trained with the same objective; they differ only in the alignment head that replaces QMORE's 80 quantum parameters.

- **Shared Classical**: a shared core Linear$(8 \to 8)$ *without bias* (64 params) plus additive per-modality bias adapters $\in \mathbb{R}^8$ ($8+8 = 16$). Replacement capacity $= 64 + 16 = 80$, a direct classical mirror of QMORE's 64 circuit + 16 adapter parameters. Total 5216.
- **Separate** $1\times$ (`SeparateBaselineStudent`): two independent encoders, each Linear$(8 \to h) + \tanh +$Linear$(h \to e)$ with $h{=}e{=}3$. Replacement $= 2\big[(8{\cdot}3{+}3) + (3{\cdot}3{+}3)\big] = 78$. Total 5214.
- **Separate** $5\times$: as above with $h{=}e{=}10$. Replacement $= 2\big[(8{\cdot}10{+}10) + (10{\cdot}10{+}10)\big] = 400$. Total 5536.
- **Shared** $5\times$: the shared mirror with its square core replaced by a widened core Linear$(8 \to 22) + \tanh +$Linear$(22 \to 8)$, keeping the 8-d embedding interface fixed. The two linear layers contribute $8{\cdot}22 + 22 + 22{\cdot}8 + 8 = 382$ parameters, and the two 8-d modality adapters contribute 16, yielding $398 \approx 5\times$ replacement-block parameters and 5534 trainable parameters in total.

Classical baseline selection. The shared baseline is also the strongest classical configuration we found at this budget: among six variants ({separate, shared} × {contrastive-only, contrastive + CLIP distillation, distillation-only}), the shared contrastive-only model was best on SVO Probes and CCIFAR and statistically indistinguishable from the CLIP-distilled variant on Flickr30k, despite the latter's access to a pretrained CLIP ViT-B/32 teacher.

C.4 Training, Ensemble, and Evaluation Protocol

Optimization. All models use Adam at learning rate 0.003 with an InfoNCE contrastive objective (temperature τ=0.07), trained for 1000 gradient steps; each step draws one minibatch of 64 pairs with replacement from the pool ($140{,}000$ image–caption pairs for Flickr30k; $8{,}000$ pairs for SVO Probes and CCIFAR).

Ensembling. Unless otherwise stated, each method uses an ensemble of $E = 10$ independently trained members. QMORE uses accuracy–diversity rank fusion with $\rho = 0.3$, rank-decay coefficient $\lambda = 0.05$, and retrieval-strength softmax temperature $\tau_\pi = 0.07$. The quantum and classical baselines use accuracy-only rank fusion.

Evaluation (std-2k). For each dataset, a fixed 500-pair calibration subset is drawn from the training pool using seed 0 and is disjoint from all evaluation data. The separate $2{,}000$-pair held-out test pool is partitioned using seed 999 into ten fixed, disjoint chunks of 200 pairs. For datasets with multiple captions per image, each chunk contains at most one caption per image. Each member ranks candidates by cosine similarity, and ensemble rankings are obtained using the corresponding rank-fusion rule. We report image→text Recall@K and the mean and median ground-truth rank as mean ± standard deviation over the ten chunks. Paired significance tests use the same ten matched chunks and are two-sided: a paired t-test and a Wilcoxon signed-rank test.

Random seeds.

- **Classical baselines:** base seed 42; the 10 ensemble members use seeds $42, 43, \ldots, 51$, controlling parameter initialization and stochastic minibatch sampling.
- **Quantum members:** training is *not* explicitly seeded per-member—member diversity arises from the OS-entropy PyTorch/NumPy initialization of the $0.01 \cdot \mathcal{U}(0, 1)$ circuit weights and $0.01 \cdot \mathcal{N}(0, 1)$ adapters together with the stochastic minibatch draws.
- **Evaluation** uses fixed chunk seeds throughout: calibration seed 0, test seed 999.

C.5 Software and Hardware

Python 3.9.6; PyTorch 2.8.0; PennyLane 0.38.0 with `pennylane-lightning` (`lightning.qubit` device for noiseless training/eval); NumPy 1.23.5; scikit-learn; `transformers` and `torchvision` (feature backbones). Noise simulation uses Qiskit 2.2.3 with `qiskit-aer`, `qiskit-ibm-runtime`, and `pennylane-qiskit`. All experiments run CPU-only on an Apple-silicon MacBook Air (Darwin 25.5).

D Noise-Model Reproducibility

This is noisy simulation, not physical-hardware execution. *All "noisy" results are produced by classical simulation (Qiskit Aer) under an IBM hardware-derived noise model. No circuit in this paper was executed on a physical IBM quantum processor.* The study characterizes robustness to a representative NISQ noise profile; it does not claim real-device results.

Backend and calibration. Qiskit's `FakeGuadalupeV2` backend Javadi-Abhari et al. (2024) supplies gate, thermal-relaxation, and readout errors extracted from hardware calibration data, applied through the PennyLane–Qiskit plugin. The noise model is built via `NoiseModel.from_backend(FakeGuadalupeV2())` from Qiskit's fake provider. `FakeGuadalupeV2` is a static snapshot of *IBM Quantum Guadalupe*, a 16-qubit Falcon-r4P processor; its device time step is $dt = 0.2222$ ns. The calibration data (gate errors, T_1/T_2, readout error) are the fixed values

baked into that fake-backend snapshot at its Qiskit release, not a live calibration. We use 8 of the 16 physical qubits.

Native gate set and transpilation. The backend basis is $\{R_Z,\ \sqrt{X}\ (\texttt{sx}),\ X,\ \texttt{id}\}$ for single-qubit gates and `cx` for two-qubit gates (plus `measure`, `reset`, and `delay`). The PennyLane–Qiskit/Aer execution path transpiles each circuit into this basis before simulation.

Qubit mapping and connectivity. Guadalupe has a heavy-hex coupling map (32 directed edges $= 16$ bidirectional CX pairs). We use 8 logical qubits. No custom initial layout or routing method is specified; the transpiler therefore selects a placement and inserts routing operations as needed to realize the ring entangler on the backend connectivity.

Noise channels. `from_backend` attaches, per gate, (i) a depolarizing error and (ii) a thermal-relaxation (T_1/T_2) error taken from the snapshot's per-qubit and per-pair calibration, plus (iii) a single-qubit readout assignment (measurement) error. No coherent/crosstalk terms beyond those in the snapshot are added.

Shots, seeds, and mitigation. The headline results use 2048 shots; noisy expectation values are shot-estimated. **No error mitigation of any kind** (no readout mitigation, no zero-noise extrapolation, and no measurement calibration) is applied; these are raw noisy estimates.