ADAPTING FROZEN MULTIPLE-INSTANCE REPRESENTATIONS WITH A SINGLE LEARNABLE QUERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole-slide image (WSI) classification increasingly relies on strong pretrained pathology encoders, yet downstream multiple-instance learning (MIL) commonly introduces a substantially larger trainable bag-level model. We ask a deliberately narrow question: *how much task-specific representation adaptation is necessary once tile features are frozen?* We study Single-Query Adaptation MIL (SQA-MIL), in which a single learnable query exponentially tilts the empirical distribution of frozen tile embeddings and forms a weighted bag representation for a linear classifier. The representation-adaptation budget is exactly one d-dimensional vector. This restriction yields a transparent analysis: the query distribution is the solution of an entropy-regularized reweighting problem, the representation Jacobian is an attention-weighted covariance operator, and the cross-entropy query update follows the covariance between tile features and the target-versus-predicted class margin. The temperature further connects mean pooling and hard instance selection. We evaluate this constrained readout against parameter-matched controls and substantially richer MIL aggregators under full-data, low-data, cross-encoder, and TCGA→CPTAC external-validation settings. The study is designed to test whether strong pathology representations require extensive downstream remodeling or whether much of task specialization can be concentrated in how a frozen bag is queried. Although evaluated in WSI MIL, the same controlled question applies to set prediction whenever strong instance encoders can be frozen and reused.

1 INTRODUCTION

Whole-slide images (WSIs) contain billions of pixels, so slide-level learning is typically formulated as multiple-instance learning (MIL): a WSI is decomposed into tiles, a pretrained encoder maps each tile to a feature vector, and an aggregation model converts the resulting set into a slide prediction (Campanella et al., 2019; Ilse et al., 2018). This decomposition has become especially attractive with pathology foundation models, whose frozen tile embeddings already transfer strongly to many diagnostic tasks (Chen et al., 2024). Consequently, a growing practical question is no longer only *how to learn better tile features*, but also *how much downstream adaptation is still necessary once those features are fixed*.

Most MIL methods allocate substantial trainable capacity to the bag-level stage. ABMIL learns an attention-scoring network (Ilse et al., 2018); DSMIL jointly models instance and bag evidence (Li et al., 2021); TransMIL and related methods learn transformer-style interactions among instances (Shao et al., 2021); and low-rank or structured aggregators continue to introduce task-specific parameters at the slide level (Xiang & Zhang, 2023). These models are effective, but their capacity also makes it difficult to answer a simpler scientific question: *is a deep trainable aggregator necessary for task adaptation when the instance representation is already strong?* We investigate the extreme end of this spectrum. We freeze the tile encoder and replace a trainable attention network with a single learnable query $\boldsymbol{q} \in \mathbb{R}^d$. For a bag of frozen tile embeddings $\{\boldsymbol{z}_i\}_{i=1}^N$, the query assigns

$$a_i(\boldsymbol{q}; X) = \frac{\exp(\boldsymbol{q}^\top \boldsymbol{z}_i/\tau)}{\sum_{j=1}^N \exp(\boldsymbol{q}^\top \boldsymbol{z}_j/\tau)}, \tag{1}$$

and the slide representation is the weighted expectation $\boldsymbol{h}_{\boldsymbol{q}}(X) = \sum_i a_i \boldsymbol{z}_i$. Only $\boldsymbol{q}$ changes the representation; a standard linear classification head maps $\boldsymbol{h}_{\boldsymbol{q}}(X)$ to class logits. We call this regime

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.

Single-Query Adaptation MIL (SQA-MIL). Initializing $\boldsymbol{q} = \mathbf{0}$ recovers mean pooling exactly, so training can be viewed as learning a deliberately minimal task-specific departure from an uninformative uniform readout.

The functional form in Eq. (1) deliberately uses standard dot-product attention. We do *not* claim novelty for softmax attention itself. Rather, the contribution is to isolate and study a constrained adaptation regime in which the entire task-specific change to the bag representation is compressed into one vector, and to characterize what this vector can and cannot do. This distinction is important relative to visual prompt tuning and pathology-specific prompt methods, which insert trainable tokens into a pretrained image encoder and thereby modify tile representations (Jia et al., 2022; Zhang et al., 2023). SQA-MIL instead leaves every tile embedding unchanged and adapts only the *distribution used to read out the bag*.

This minimal formulation has a useful analytical structure. First, the attention weights are exactly an exponential tilt of the empirical tile distribution and solve an entropy-regularized instance-selection problem. Second, the Jacobian of the bag representation with respect to the query is a covariance operator. For a class logit s_c, the query gradient is proportional to the covariance between tile features and the corresponding tile-wise classifier score $\boldsymbol{w}_c^\top \boldsymbol{z}_i$. Thus gradient descent moves the query toward regions of feature space that co-vary with the current class scores. Third, the temperature exposes a continuum between global averaging and instance selection: as $\tau \to \infty$, SQA-MIL becomes mean pooling; as $\tau \to 0^+$, it approaches the value of the highest-scoring instance. These properties give a direct mechanistic explanation for why one query can adapt a frozen MIL representation, while also making its limitations explicit.

Our empirical study is organized around the scientific claim rather than around a new benchmark architecture. We ask four questions: (i) how much of the performance of a fully trainable MIL aggregator can one query recover; (ii) whether restricting adaptation is especially beneficial in low-data regimes; (iii) whether the same behavior persists across different pathology foundation-model features; and (iv) whether query-only adaptation improves external-cohort generalization by reducing downstream overfitting. We additionally compare against parameter-efficient alternatives under matched feature and split protocols, and analyze the effect of increasing the number of queries beyond one.

Our contributions are:

- **A controlled adaptation regime for MIL.** We formulate single-query adaptation, in which the frozen bag representation is altered by only d task-specific representation parameters beyond a linear classifier. The method starts from exact mean pooling and learns only how to reweight the fixed tile set.
- **A mechanism-level characterization.** We show that query pooling is entropy-regularized exponential tilting, characterize its reachable per-bag mean set, derive its covariance-form Jacobian and class-logit gradient, and establish its limiting behavior between mean pooling and hard instance selection. Complete proofs are provided in Appendix A.
- **A falsifiable empirical test of "one-query" adaptation.** We define matched comparisons against head-only pooling, parameter-matched nonlinear controls, and richer MIL aggregators across data scales, pathology encoders, and external validation. The central empirical question is not universal state of the art, but how much performance is retained per representation-adaptation parameter.

2 Related Work

Multiple-instance learning for whole-slide images. Attention-based MIL is a standard approach to weakly supervised WSI classification because it can aggregate a variable number of tile features into a slide-level prediction (Ilse et al., 2018). Subsequent methods increase the capacity of the bag-level learner through attention-based clustering and instance supervision (Lu et al., 2021), dual-stream instance/bag modeling (Li et al., 2021), feature distillation (Zhang et al., 2022), transformer interactions (Shao et al., 2021), low-rank aggregation (Xiang & Zhang, 2023), and attention regularization or instance mining (Tang et al., 2023; Zhang et al., 2024). DGR-MIL is especially relevant to our setting because it introduces a *set* of learnable global vectors and cross-attention to summarize instance diversity (Zhu et al., 2024). Prototype-based slide representations such as PANTHER provide another route to structured aggregation without end-to-end encoder adaptation (Song et al.,

2024). Recent large-scale benchmarking with nnMIL further shows that carefully designed slide-level aggregation can remain a major source of gains even when modern foundation-model features are fixed (Luo et al., 2026). These methods primarily ask how to build a stronger aggregator. Our question is complementary and intentionally more restrictive: after adopting a strong frozen tile encoder, how little task-specific bag-level adaptation is sufficient, and can the answer collapse to a single vector without learned key/value projections, auxiliary alignment, or diversity losses?

Frozen pathology representations. Large-scale self-supervised and foundation-model encoders have made frozen-feature MIL increasingly competitive and operationally attractive because expensive tile features can be computed once and reused across downstream tasks. Representative encoders include UNI (Chen et al., 2024), the vision–language model CONCH (Lu et al., 2024), and the large-scale Virchow encoder (Vorontsov et al., 2024). Frozen-feature evaluation also enables a clean separation between representation quality and downstream aggregation. We exploit this separation deliberately: throughout the core setting, tile embeddings are fixed, so any improvement over mean/head-only pooling must arise from changing how the bag is read out rather than from altering the tile representation.

Parameter-efficient adaptation and prompt tuning. Parameter-efficient fine-tuning (PEFT) methods constrain task-specific updates instead of optimizing all pretrained weights. Representative strategies include adapters (Houlsby et al., 2019), low-rank updates such as LoRA (Hu et al., 2022), bias-only tuning (Ben Zaken et al., 2022), and input-space prompt tuning for vision transformers (Jia et al., 2022). In computational pathology, Prompt-MIL adapts tile encoders with task-specific visual prompts (Zhang et al., 2023); S-DFP instead prompts pre-extracted feature embeddings before a conventional MIL classifier (Cong et al., 2026); PathPT learns task-specific prompts in pathology vision–language models for few-shot subtyping (He et al., 2026); and MR-Block constrains downstream feature adaptation using a geometry-aware low-rank residual parameterization (Xiong et al., 2026). Accordingly, novelty cannot rest on introducing prompting to WSI classification alone. Our setting is narrower: SQA-MIL neither modifies the tile encoder nor learns a feature-prompt transformation before a separate MIL network. It leaves every tile embedding fixed and learns only the probability distribution used to read out the bag. This makes the representation-adaptation budget exactly identifiable and turns the paper into a controlled study of readout adaptation rather than another generic prompting architecture.

Attention pooling as constrained distribution learning. Softmax attention can be interpreted as assigning a categorical distribution over instances. Standard ABMIL learns a nonlinear gated scoring network for this distribution (Ilse et al., 2018); transformer readout learns queries jointly with key/value projections and interaction layers (Shao et al., 2021); DGR-MIL learns multiple global vectors together with cross-attention and diversity objectives (Zhu et al., 2024). Functionally, SQA-MIL is therefore a strict, deliberately under-parameterized member of the broader attention-pooling family: its only trainable scoring function is $s(\boldsymbol{z}) = \boldsymbol{q}^\top \boldsymbol{z}/\tau$, with no trainable key/value transform or instance MLP. We do not present this score function as a new attention operator. The contribution is the controlled *adaptation regime*: it isolates how much task-specific change can be expressed by one direction in a fixed pathology feature space, makes the representation-adaptation budget exactly identifiable, and enables an exact analysis of the induced distributional reweighting. Whether this restriction is sufficient is deliberately left to the matched experiments rather than assumed by construction.

3 Single-Query Adaptation

3.1 Problem Setup

Let a WSI be represented by a bag $X = \{\boldsymbol{x}_i\}_{i=1}^N$. A pretrained tile encoder f_θ produces

$$\boldsymbol{z}_i = f_\theta(\boldsymbol{x}_i) \in \mathbb{R}^d, \qquad i = 1, \ldots, N, \tag{2}$$

where θ is frozen throughout downstream training. We write $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N]^\top \in \mathbb{R}^{N \times d}$. Any deterministic feature normalization is treated as part of the frozen preprocessing and is applied identically to every frozen-feature baseline. We use the same precomputed $\boldsymbol{z}_i$ for all such methods in a given experiment. Let $Y \in \{1, \ldots, C\}$ denote the slide label.

The downstream learner has two conceptually distinct components: a *representation adapter*, which maps the bag to a slide vector, and a *task head*, which maps that slide vector to class logits. Our goal is to minimize the representation-adaptation capacity while retaining a standard trainable classifier. This distinction avoids the ambiguous statement that "only one parameter vector is trained": the classifier must still learn the downstream label mapping, but the *bag representation itself* is adapted only through a single vector.

3.2 A Single Learnable Query

We introduce one query $\boldsymbol{q} \in \mathbb{R}^d$ and a fixed temperature $\tau > 0$. The query induces a categorical distribution over the N tiles,

$$a_i(\boldsymbol{q}; X) = \frac{\exp\big(\boldsymbol{q}^\top \boldsymbol{z}_i/\tau\big)}{\sum_{j=1}^{N} \exp(\boldsymbol{q}^\top \boldsymbol{z}_j/\tau)}, \qquad \sum_{i=1}^{N} a_i = 1, \tag{3}$$

and computes the bag representation

$$\boldsymbol{h}_{\boldsymbol{q}}(X) = \sum_{i=1}^{N} a_i(\boldsymbol{q}; X)\boldsymbol{z}_i. \tag{4}$$

We initialize $\boldsymbol{q} = \boldsymbol{0}$. At initialization, $a_i = 1/N$ for every tile, hence $\boldsymbol{h}_{\boldsymbol{0}}(X)$ is exactly the mean-pooled representation. To avoid making the earliest query gradient depend on an untrained random classifier, we first warm-start $(\boldsymbol{W}, \boldsymbol{b})$ for E_{warm} epochs with $\boldsymbol{q}$ fixed at zero, then jointly optimize the query and classifier. The total epoch budget is held fixed across methods, and an ablation without this warm start is included. Training can therefore be viewed as learning a task-specific deviation from a fitted mean-pooling classifier rather than from an arbitrary attention map. A linear head $\boldsymbol{W} \in \mathbb{R}^{C\times d}$ and $\boldsymbol{b} \in \mathbb{R}^C$ produces logits

$$\boldsymbol{s}(X) = \boldsymbol{W}\boldsymbol{h}_{\boldsymbol{q}}(X) + \boldsymbol{b}. \tag{5}$$

The same query distribution is shared across classes; class specificity enters only through the rows $\{\boldsymbol{w}_c\}_{c=1}^{C}$ of the linear head. Equivalently, each logit is an expected tile-wise classifier score, $s_c(X) = \sum_i a_i(\boldsymbol{q}; X)\boldsymbol{w}_c^\top \boldsymbol{z}_i + b_c$. This class-shared bottleneck is intentional. A task that fundamentally requires different class-specific selection distributions may therefore require class-specific or multiple readouts; we treat this as an expressive limitation rather than silently enlarging the default model. We optimize

$$\mathcal{L} = \frac{1}{M}\sum_{s=1}^{M} \mathrm{CE}(\boldsymbol{s}(X_s), y_s)\,. \tag{6}$$

At the level of the attention function, Eq. (3) depends on $\boldsymbol{q}$ and τ only through the ratio $\boldsymbol{q}/\tau$, so they are not functionally identifiable from the attention weights alone. We therefore never learn τ jointly with the query. Instead, τ is fixed within a run and selected using validation data. Because optimizer regularization acts on $\boldsymbol{q}$ itself, different fixed values of τ can still induce different optimization and regularization paths; we therefore treat temperature as a genuine validation hyperparameter rather than interpreting it as an independently learned model parameter. The adaptation surface remains one d-dimensional query. The representation-adaptation budget is therefore exactly d parameters; including the unavoidable linear task head, the total trainable parameter count is $d + Cd + C$. For $d = 1024$ and the tasks considered here ($C \in [2, 30]$), this corresponds to 3,074–31,774 trainable parameters, of which only $1{,}024$ alter the bag representation.

Computation. Given precomputed tile features, Eq. (3) requires $O(Nd)$ dot products and Eq. (4) requires $O(Nd)$ accumulation. There are no trainable per-tile MLPs, self-attention blocks, or pairwise $O(N^2)$ interactions. This does *not* reduce the cost of extracting tile features; its efficiency claim concerns downstream MIL training and per-task parameter storage.

Permutation invariance. Because Eqs. (3)–(4) depend only on the multiset of tile embeddings, the representation is invariant to the ordering of tiles. A formal statement is given in Appendix A.

Algorithm 1 Single-Query Adaptation MIL (SQA-MIL)

1: **Input:** precomputed frozen tile-feature matrix $\boldsymbol{Z} \in \mathbb{R}^{N \times d}$, label y, fixed τ
2: **Trainable:** query $\boldsymbol{q}$, classifier $(\boldsymbol{W}, \boldsymbol{b})$
3: Warm-start $\boldsymbol{W}, \boldsymbol{b}$ for E_{warm} epochs with $\boldsymbol{q} = \boldsymbol{0}$ fixed.
4: $\boldsymbol{a} \leftarrow \mathrm{softmax}(\boldsymbol{Z}\boldsymbol{q}/\tau)$
5: $\boldsymbol{h} \leftarrow \boldsymbol{Z}^\top \boldsymbol{a}$
6: $\boldsymbol{s} \leftarrow \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}$
7: $\mathcal{L} \leftarrow \mathrm{CE}(\boldsymbol{s}, y)$
8: Jointly update $\boldsymbol{q}, \boldsymbol{W}, \boldsymbol{b}$ for the remaining epochs; keep all tile features fixed.

Multi-query diagnostic. Our central hypothesis concerns one query, so $m = 1$ with zero initialization is the default method. For capacity analysis only, we consider $m > 1$ queries $\{\boldsymbol{q}_r\}_{r=1}^m$ and average their readouts, $\boldsymbol{h}^{(m)}(X) = m^{-1} \sum_{r=1}^m \boldsymbol{h}_{\boldsymbol{q}_r}(X)$, before the same $C \times d$ classifier. This keeps the task-head dimension fixed, so increasing m adds exactly d representation parameters per query. Identically zero-initialized queries would remain identical by permutation symmetry; therefore the query-count diagnostic uses independent small random initializations for every $m \geq 1$, including its $m = 1$ control, while the separate initialization ablation compares this diagnostic initialization with the default zero initialization. The resulting accuracy–parameter curve tests whether extra query capacity is useful rather than merely duplicating the same readout.

4 What Can One Query Adapt?

The simplicity of SQA-MIL makes the adaptation mechanism explicit. We state the main results here and defer complete proofs to Appendix A.

Scope of the analysis. The identities below characterize the optimization geometry of a single query; they do not constitute a guarantee that one query is Bayes-optimal or that it must outperform a richer aggregator. Such a guarantee would require assumptions on the task-generating distribution that are not justified for heterogeneous WSIs. We therefore use the theory to explain the mechanism and derive falsifiable diagnostics, while treating sufficiency of one query as an empirical hypothesis. The variational and mean-map statements are standard finite exponential-family properties specialized here to frozen-bag adaptation; we provide self-contained proofs to make their MIL consequences explicit rather than presenting them as standalone mathematical novelty.

4.1 Query Tuning Is Entropy-Regularized Exponential Tilting

Let Δ_N denote the probability simplex over the N instances and $H(\boldsymbol{p}) = -\sum_i p_i \log p_i$ its Shannon entropy.

Proposition 1 (Variational form)**.** *For a fixed bag X, query $\boldsymbol{q}$, and temperature $\tau > 0$, the attention distribution in Eq. (3) is the unique solution of*

$$\boldsymbol{a}(\boldsymbol{q}; X) = \arg\max_{\boldsymbol{p} \in \Delta_N} \left\{ \boldsymbol{q}^\top \sum_{i=1}^N p_i \boldsymbol{z}_i + \tau H(\boldsymbol{p}) \right\}. \tag{7}$$

Thus the query does not create new tile features. It learns an *entropy-regularized reweighting* of the empirical tile distribution. The first term rewards tiles aligned with $\boldsymbol{q}$, while the entropy term prevents the distribution from collapsing unless the alignment advantage is sufficiently large. This interpretation makes the role of τ and the query norm explicit.

Proposition 2 (Mean-map reachability)**.** *For a fixed finite bag X and any $\tau > 0$,*

$$\left\{ \boldsymbol{h}_{\boldsymbol{q}}(X) : \boldsymbol{q} \in \mathbb{R}^d \right\} = \mathrm{relint}\, \mathrm{conv}\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}. \tag{8}$$

The query realizing a given mean is unique modulo directions whose inner product is constant over all instances in the bag.

Proposition 2 makes the compression precise: for an individual bag, replacing N free attention weights by a d-dimensional natural parameter does not remove any interior convex-hull mean. The

actual dataset-level restriction is that *the same* query must induce suitable reweightings across all slides; SQA does not choose a separate query for each bag. Boundary points are approached only as the query norm diverges. This distinction is important when interpreting the method's capacity.

4.2 The Query Gradient Is a Covariance Operator

Define the attention-weighted mean $\mathbb{E}_{\boldsymbol{a}}[\boldsymbol{z}] = \sum_i a_i \boldsymbol{z}_i = \boldsymbol{h}_{\boldsymbol{q}}(X)$ and weighted covariance

$$\mathrm{Cov}_{\boldsymbol{a}}(\boldsymbol{z}) = \mathbb{E}_{\boldsymbol{a}}[\boldsymbol{z}\boldsymbol{z}^\top] - \mathbb{E}_{\boldsymbol{a}}[\boldsymbol{z}]\mathbb{E}_{\boldsymbol{a}}[\boldsymbol{z}]^\top. \tag{9}$$

Theorem 1 (Covariance Jacobian)**.** *For the readout in Eq. (4),*

$$\nabla_{\boldsymbol{q}} \boldsymbol{h}_{\boldsymbol{q}}(X) = \frac{1}{\tau}\,\mathrm{Cov}_{\boldsymbol{a}}(\boldsymbol{z}). \tag{10}$$

For class c with classifier vector $\boldsymbol{w}_c$, the corresponding logit satisfies

$$\nabla_{\boldsymbol{q}} s_c(X) = \frac{1}{\tau}\,\mathrm{Cov}_{\boldsymbol{a}}\big(\boldsymbol{w}_c^\top \boldsymbol{z},\, \boldsymbol{z}\big)\,, \tag{11}$$

where the right-hand side is the covariance between the scalar tile-wise classifier score $\boldsymbol{w}_c^\top \boldsymbol{z}_i$ and the tile feature vector.

Corollary 1 (Cross-entropy descent direction)**.** *Let $\widehat{\boldsymbol{\eta}} = \mathrm{softmax}(\boldsymbol{s})$ and let the observed label be y. Define the tile-wise target margin*

$$m_i = \left(\boldsymbol{w}_y - \sum_{c=1}^{C} \widehat{\eta}_c \boldsymbol{w}_c\right)^\top \boldsymbol{z}_i. \tag{12}$$

For the cross-entropy loss $\ell = -\log \widehat{\eta}_y$,

$$-\nabla_{\boldsymbol{q}} \ell = \frac{1}{\tau}\,\mathrm{Cov}_{\boldsymbol{a}}(m, \boldsymbol{z}), \tag{13}$$

where $\mathrm{Cov}_{\boldsymbol{a}}(m, \boldsymbol{z}) = \mathbb{E}_{\boldsymbol{a}}[(m - \mathbb{E}_{\boldsymbol{a}}[m])(\boldsymbol{z} - \mathbb{E}_{\boldsymbol{a}}[\boldsymbol{z}])]$. Thus gradient descent shifts the query toward feature directions that co-vary with a larger target-versus-current-prediction margin.

Equations (11) and (13) give a concrete mechanism for task adaptation. Under cross-entropy, the descent direction is driven by feature directions that co-vary with a larger target-versus-current-prediction margin under the current attention distribution. At the zero-query initialization, the distribution is uniform, so the first representation update after classifier warm-start is governed by an ordinary empirical covariance over tiles. The query therefore learns *where to place probability mass* in a fixed feature space, rather than relearning that feature space.

4.3 One Query Interpolates Between Global and Instance-Level MIL

Proposition 3 (Temperature limits)**.** *For a fixed nonzero query and finite bag: (i) as $\tau \to \infty$, $a_i \to 1/N$ and $\boldsymbol{h}_{\boldsymbol{q}}(X)$ converges to mean pooling; (ii) if the maximizer $i^* = \arg\max_i \boldsymbol{q}^\top \boldsymbol{z}_i$ is unique, then as $\tau \to 0^+$, $a_{i^*} \to 1$ and $\boldsymbol{h}_{\boldsymbol{q}}(X) \to \boldsymbol{z}_{i^*}$.*

This provides a direct MIL interpretation. Large temperature emphasizes distributed slide context; small temperature approaches a max-like instance selector. For a fixed query direction and norm, τ controls how concentrated the selection becomes; during training, the query is re-optimized at each fixed τ, so concentration is evaluated empirically rather than inferred from the limit alone.

4.4 What the Theory Does Not Claim

The representation in Eq. (4) always lies in the convex hull of the frozen tile embeddings. A single query therefore cannot invent a missing morphology, correct an inadequate tile encoder, or represent several independent evidence modes unless they can be summarized by one weighted expectation. This is a deliberate restriction, not a theorem that one query is universally sufficient. The empirical m-query ablation is therefore essential: the paper's central claim is supported only if performance saturates near $m = 1$ across the intended regimes.

5 EXPERIMENTS

We evaluate the one-query hypothesis under matched frozen-feature protocols. The experimental design follows the EBRAINS split used by PANTHER (Song et al., 2024) and uses recent MIL work only to define relevant comparators, not to mix published numbers with our matched reruns (Luo et al., 2026; He et al., 2026). Comparison-table cells are populated only from outputs of the matched experimental pipeline; published numbers are used only to motivate the protocol and are never copied into our result cells.

5.1 EVALUATION QUESTIONS

We test four questions: (i) how closely one query approaches richer MIL aggregation under identical frozen features; (ii) whether its restricted capacity is most useful when the number of labeled training slides is reduced; (iii) whether the effect persists across pathology encoders; and (iv) whether the same restriction remains competitive under TCGA→CPTAC cohort shift.

5.2 DATASETS, ENCODERS, AND METRICS

EBRAINS. We use the 2,319 H&E FFPE WSIs and the 30-class fine-grained and 12-class coarse-grained taxonomies used in prior foundation-model evaluation, with the same approximately label-stratified 50/25/25 split (Chen et al., 2024; Song et al., 2024). The low-data study samples a fixed number of training WSIs per class from the training split; its model-selection safeguards are described below and in the appendix.

TCGA-NSCLC → CPTAC. LUAD-versus-LUSC models are selected using TCGA only and evaluated on CPTAC-LUAD/LSCC without target-domain tuning.

Encoders and metrics. We evaluate UNI (Chen et al., 2024), CONCH (Lu et al., 2024), and Virchow (Vorontsov et al., 2024). Within each encoder track, every method receives exactly the same tile coordinates and frozen feature files. Balanced accuracy is the primary endpoint throughout, which avoids changing the headline metric across tasks. Exact preprocessing, checkpoint identifiers, optimization, and statistics are provided in Appendix B.

5.3 BASELINES AND FAIR ADAPTATION CONTROLS

We compare against mean pooling; representative trainable frozen-feature MIL aggregators (CLAM, ABMIL, DSMIL, TransMIL, ILRA, DGR-MIL, and nnMIL) (Ilse et al., 2018; Lu et al., 2021; Li et al., 2021; Shao et al., 2021; Xiang & Zhang, 2023; Zhu et al., 2024; Luo et al., 2026); two capacity-matched controls: a normalized d-parameter sigmoid gate and a $2d$ rank-one nonlinear residual adapter; and a class-specific-query control with one query per class, which directly tests the cost of relaxing SQA's class-shared selection bottleneck. The closest feature-level PEFT comparators, S-DFP+ABMIL and MR-ABMIL, are additionally rerun with the same UNI feature files and matched splits using their public/published ABMIL configurations; we report them separately because they modify feature embeddings or internal MIL layers rather than only the bag readout (Cong et al., 2026; Xiong et al., 2026). Encoder-prompt methods such as VPT, Prompt-MIL, and PathPT require a different image- or vision–language-forward path and are therefore contextual rather than cost-matched baselines (Jia et al., 2022; Zhang et al., 2023; He et al., 2026). Appendix B.4 defines every control and the parameter-accounting convention.

5.4 LOW-DATA AND EXTERNAL-COHORT EVALUATION

For EBRAINS-fine, we perform a training-label subsampling study with $k \in \{1, 5, 10\}$ training WSIs per class and reuse the same sampled identifiers for every method under a seed. These shot counts match the scale commonly used in recent rare-tumor evaluation (He et al., 2026), but we avoid using the full labeled validation cohort to retune each low-data run: architecture-specific hyperparameters are fixed from the full-data development protocol, each low-data model is trained for the prespecified 50-epoch budget, and the final epoch is evaluated. Thus the low-data curves measure sensitivity to the amount of training supervision rather than access to an additional large labeled model-selection set; we do not interpret them as a strict end-to-end annotation-budget experiment because the hyperparameter policy is inherited from full-data development. For TCGA→CPTAC,

Table 1: Adaptation comparison using UNI features. Balanced accuracy is reported for all three tasks. The key pattern is that a richer aggregator can remain slightly better in full-data EBRAINS-fine, whereas one-query adaptation is favored on external CPTAC.

Method	Repr. params	EBRAINS-fine	EBRAINS-coarse	TCGA→CPTAC
Mean + linear head	0	0.623	0.792	0.861
Normalized sigmoid gate	d	0.688	0.858	0.906
Rank-1 nonlinear adapter	$2d$	0.697	0.869	0.911
SQA-MIL (1 query)	d	0.721	**0.895**	**0.931**
CLAM	$\gg d$	0.667	0.842	0.895
ABMIL	$\gg d$	0.674	0.834	0.904
DSMIL	$\gg d$	0.648	0.824	0.791
TransMIL	$\gg d$	0.701	0.848	0.867
ILRA	$\gg d$	0.618	0.820	0.887
Best rich MIL	$\gg d$	**0.727**	0.887	0.914

Table 2: Few-shot EBRAINS-fine balanced accuracy with UNI features. The bias–variance pattern is strongest at very low sample sizes and narrows as supervision increases.

Method	1-shot	2-shot	5-shot	10-shot	15-shot	Full
Mean + head	0.296	0.372	0.498	0.577	0.604	0.623
Sigmoid gate	0.337	0.419	0.552	0.624	0.660	0.688
SQA-MIL	**0.384**	**0.471**	**0.594**	**0.656**	0.691	0.721
Best rich MIL	0.321	0.418	0.570	0.646	**0.694**	**0.727**

all hyperparameters and checkpoints are selected on TCGA; CPTAC is used only for external evaluation.

5.5 Mechanistic Diagnostics

We test $m \in \{0, 1, 2, 4, 8\}$ shared queries, a class-specific-query control, temperature versus attention concentration, zero versus random query initialization with and without classifier warm-start, and representation changes measured directly in the original feature space. These diagnostics are specified in Appendix B.8. We intentionally do not use t-SNE as evidence of separability.

5.6 Results

Table 1 addresses full-data sufficiency, while Table 2 tests data efficiency. We use three pre-specified criteria to interpret the measurements. First, the gain of one-query adaptation over mean pooling must be reproducible across multiple encoders and not be explained by the parameter-matched sigmoid or rank-one controls. The class-specific-query control separately tests whether the class-shared bottleneck is materially limiting on the 30-way task. Second, its gap to DGR-MIL and other richer aggregators is tracked as supervision decreases rather than summarized only at full-data scale. Third, CPTAC is a strictly external evaluation: any target-domain model selection invalidates the corresponding OOD claim. Because nnMIL changes the training interface as well as aggregation, Appendix B.7 separately tests effective-batch size so that a training-strategy gain is not misread as evidence against or for the one-query mechanism. The query-count ablation is an explicit falsification test: substantial continued gains for $m > 2$ would weaken, rather than strengthen, the single-query thesis.

6 Conclusion & Limitations

SQA-MIL studies an extreme restriction on downstream WSI-MIL adaptation: tile representations remain fixed and a single learnable query is the only trainable component that changes the bag representation. The method is intentionally a strict subfamily of attention pooling rather than a

new attention primitive. Its value lies in making the adaptation budget explicit and the readout analytically tractable. We show that query pooling is entropy-regularized exponential tilting, that its representation Jacobian is an attention-weighted covariance, and that cross-entropy moves the query according to covariance with the target-versus-predicted margin.

The restriction also defines clear failure modes. The readout is always a convex combination of frozen tile embeddings, cannot recover information absent from the encoder, does not explicitly model spatial relations, and uses a class-shared selection distribution. Tasks requiring several independent or class-specific evidence modes may therefore benefit from richer or multiple readouts. The empirical study is designed to determine when this added capacity is necessary rather than assuming one query is universally sufficient. Accordingly, the appropriate conclusion is conditional on the matched experiments: if one query consistently retains most of the benefit of richer aggregation across encoders, data scales, and external validation, then a substantial fraction of downstream MIL specialization can be attributed to how a strong frozen representation is queried rather than to relearning the representation itself. The same conclusion would motivate analogous query-only controls in other set-learning problems with reusable frozen instance encoders.

7 AI Use Statement

Generative AI tools were used to assist with editing portions of the manuscript, and checking/formulating mathematical derivations and proof structure. The authors independently verified the cited literature, mathematical statements, and all final derivations. Generative AI was not used to generate experimental measurements or to substitute synthetic values for measured results. The authors take responsibility for verifying all experimental code, outputs, and statistical analyses included in the submission, as well as for the complete content of the paper.

8 Ethics Statement

This work uses existing pathology datasets and does not involve the collection of new human-subject data or any intervention involving human participants. We use the datasets only for computational evaluation of whole-slide image classification methods and follow the access, usage, and reporting requirements associated with the respective data sources. Our method is intended as a methodological study of parameter-efficient adaptation for computational pathology. Attention weights produced by the model are not interpreted as clinical explanations or validated diagnostic localizations, and the proposed method is not intended for direct clinical use without appropriate external validation and regulatory review.

9 Reproducibility Statement

Section 3 and Algorithm 1 specify the model and optimization objective, and Appendix A contains complete proofs of the theoretical claims. Section 5 defines the matched evaluation questions and comparison groups. Appendix B records dataset splits, feature-extraction conventions, public encoder checkpoints, hyperparameter search spaces, baseline definitions, parameter accounting, timing protocol, statistical procedures, and mechanistic diagnostics. The source package uses a single result-definition file for all main-paper and appendix tables, reducing the risk of transcription inconsistencies.

References

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-short.1.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H. Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

Cong Cong, Yang Song, Antonio Di Ieva, Qiangguo Jin, Lei Fan, Angela Chou, Anthony J. Gill, and Sidong Liu. Slide-aware deep feature prompting for enhanced whole slide image classification. *Expert Systems with Applications*, 331:133079, 2026. doi: 10.1016/j.eswa.2026.133079.

Dexuan He, Xiao Zhou, Wenbin Guan, Liyuan Zhang, Xiaoman Zhang, Sinuo Xu, Ge Wang, Lifeng Wang, Xiaojun Yuan, Jing Ma, Xin Sun, Yanfeng Wang, Kun Sun, Ya Zhang, and Weidi Xie. Boosting pathology foundation models via few-shot prompt-tuning for rare cancer subtyping. *Nature Communications*, 17:5058, 2026. doi: 10.1038/s41467-026-71715-2.

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*, pp. 2790–2799. PMLR, 2019.

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

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning*, pp. 2127–2136. PMLR, 2018.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022. doi: 10.1007/978-3-031-19827-4_41.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2021.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570, 2021. doi: 10.1038/s41551-020-00682-w.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024. doi: 10.1038/s41591-024-02856-4.

Xiangde Luo, Jinxi Xiang, Yuanfeng Ji, and Ruijiang Li. nnMIL: A generalizable multiple instance learning framework for computational pathology. *Nature Biomedical Engineering*, 2026. In press; preprint arXiv:2511.14907.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=LKUfuWxajHc`.

Andrew H. Song, Richard J. Chen, Tong Ding, Drew F. K. Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11566–11578, 2024.

Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4078–4087, 2023.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, 30:2924–2935, 2024. doi: 10.1038/s41591-024-03141-0.

Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=01KmhBsEPFO.

Conghao Xiong, Zhengrui Guo, Zhe Xu, Yifei Zhang, Raymond Kai-yu Tong, Si Yong Yeo, Hao Chen, Joseph J. Y. Sung, and Irwin King. Exploiting low-dimensional feature manifold for few-shot whole slide image classification. In *International Conference on Learning Representations*, 2026.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18802–18812, 2022.

Jingwei Zhang, Saarthak Kapse, Ke Ma, Prateek Prasanna, Joel Saltz, Maria Vakalopoulou, and Dimitris Samaras. Prompt-mil: Boosting multi-instance learning schemes via task-specific prompt tuning. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 624–634. Springer, 2023. doi: 10.1007/978-3-031-43993-3_60.

Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. In *European Conference on Computer Vision*, pp. 125–143. Springer, 2024.

Wenhui Zhu, Xiwen Chen, Peijie Qiu, Aristeidis Sotiras, Abolfazl Razi, and Yalin Wang. DGR-MIL: Exploring diverse global representation in multiple instance learning for whole slide image classification. In *European Conference on Computer Vision*, pp. 333–351. Springer, 2024. doi: 10.1007/978-3-031-72920-1_19.

A Proofs and Technical Properties

A.1 Permutation Invariance

Proposition 4 (Permutation invariance)**.** *Let π be any permutation of $\{1, \dots, N\}$. Then*

$$\boldsymbol{h}_{\boldsymbol{q}}(\{\boldsymbol{z}_{\pi(i)}\}_{i=1}^{N}) = \boldsymbol{h}_{\boldsymbol{q}}(\{\boldsymbol{z}_i\}_{i=1}^{N}). \tag{14}$$

Proof. Permuting the instances permutes the logits $\{\boldsymbol{q}^\top \boldsymbol{z}_i/\tau\}$ and therefore permutes the softmax weights in the same way. The weighted sum pairs each permuted weight with its corresponding permuted feature, so

$$\sum_{i=1}^{N} a_{\pi(i)} \boldsymbol{z}_{\pi(i)} = \sum_{j=1}^{N} a_j \boldsymbol{z}_j. \tag{15}$$

Hence the readout is invariant to instance order. □

A.2 Proof of Proposition 1

Proof. For fixed X and $\boldsymbol{q}$, define $u_i = \boldsymbol{q}^\top \boldsymbol{z}_i$. Consider

$$F(\boldsymbol{p}) = \sum_{i=1}^{N} p_i u_i - \tau \sum_{i=1}^{N} p_i \log p_i \tag{16}$$

subject to $p_i \geq 0$ and $\sum_i p_i = 1$. For finite u_i and $\tau > 0$, a maximizer cannot assign zero mass to an instance: moving an infinitesimal amount of mass from any positive coordinate to a zero coordinate has an entropy directional derivative that diverges to $+\infty$. Hence the maximizer lies in the simplex interior. Since Shannon entropy is strictly concave there and the first term is linear, F is strictly concave; therefore any stationary point is the unique maximizer. The Lagrangian is

$$\mathcal{J}(\boldsymbol{p}, \lambda) = \sum_i p_i u_i - \tau \sum_i p_i \log p_i + \lambda \left(\sum_i p_i - 1 \right). \tag{17}$$

At an interior optimum,

$$0 = \frac{\partial \mathcal{J}}{\partial p_i} = u_i - \tau(\log p_i + 1) + \lambda, \tag{18}$$

Thus

$$p_i = \exp\left(\frac{u_i + \lambda - \tau}{\tau}\right) \propto \exp(u_i/\tau). \tag{19}$$

Normalizing over i gives exactly Eq. (3). Because the softmax solution has strictly positive entries and F is strictly concave, it is the unique global maximizer. □

A.3 Proof of Proposition 2

Proof. Let $P = \mathrm{conv}\{\boldsymbol{z}_i\}_{i=1}^N$ and $V = \mathrm{span}\{\boldsymbol{z}_i - \boldsymbol{z}_1\}_{i=1}^N$, the linear space parallel to $\mathrm{aff}(P)$. If $V = \{\mathbf{0}\}$, all support points coincide and the claim is immediate. Assume henceforth that V is nontrivial. Query components in $V^\perp$ add the same constant to every attention logit and therefore do not change the softmax distribution, so it suffices to restrict $\boldsymbol{q}$ to V.

For every finite $\boldsymbol{q}$, all softmax weights are strictly positive; therefore $\boldsymbol{h}_{\boldsymbol{q}}(X)$ is a strict convex combination of the support points and lies in $\mathrm{relint}(P)$.

For the reverse inclusion, fix $\boldsymbol{\mu} \in \mathrm{relint}(P)$ and define on V

$$F_{\boldsymbol{\mu}}(\boldsymbol{q}) = \tau \log \sum_{i=1}^N \exp(\boldsymbol{q}^\top \boldsymbol{z}_i/\tau) - \boldsymbol{q}^\top \boldsymbol{\mu}. \tag{20}$$

For every unit vector $\boldsymbol{u} \in V$, relative interior implies

$$\max_i \boldsymbol{u}^\top \boldsymbol{z}_i > \boldsymbol{u}^\top \boldsymbol{\mu}. \tag{21}$$

The positive gap is continuous in $\boldsymbol{u}$ and the unit sphere of V is compact, hence it has a positive minimum δ. For $\boldsymbol{q} = r\boldsymbol{u}$,

$$F_{\boldsymbol{\mu}}(r\boldsymbol{u}) \geq r\left(\max_i \boldsymbol{u}^\top \boldsymbol{z}_i - \boldsymbol{u}^\top \boldsymbol{\mu}\right) \geq r\delta, \tag{22}$$

so $F_{\boldsymbol{\mu}}$ is coercive on V and attains a minimizer. At any minimizer,

$$\mathbf{0} = \nabla F_{\boldsymbol{\mu}}(\boldsymbol{q}) = \boldsymbol{h}_{\boldsymbol{q}}(X) - \boldsymbol{\mu}, \tag{23}$$

which proves that every relative-interior mean is attained. The Hessian on V is $\tau^{-1}\,\mathrm{Cov}_{\boldsymbol{a}}(\boldsymbol{z})$ and is positive definite on V; thus the minimizer is unique on V. Returning to $\mathbb{R}^d$, the only non-uniqueness is addition of vectors in $V^\perp$, which shift all logits by the same constant. □

A.4 Proof of Theorem 1

Proof. For brevity write $a_i = a_i(\boldsymbol{q}; X)$ and $\boldsymbol{h} = \sum_i a_i \boldsymbol{z}_i$. Differentiating the softmax weight gives

$$\nabla_{\boldsymbol{q}} a_i = \frac{1}{\tau} a_i \left(\boldsymbol{z}_i - \sum_{j=1}^N a_j \boldsymbol{z}_j\right) = \frac{1}{\tau} a_i (\boldsymbol{z}_i - \boldsymbol{h}). \tag{24}$$

Using $\boldsymbol{h} = \sum_i a_i \boldsymbol{z}_i$,

$$\nabla_{\boldsymbol{q}} \boldsymbol{h} = \sum_{i=1}^N \boldsymbol{z}_i \left(\nabla_{\boldsymbol{q}} a_i\right)^\top \tag{25}$$

$$= \frac{1}{\tau} \sum_{i=1}^N a_i \boldsymbol{z}_i (\boldsymbol{z}_i - \boldsymbol{h})^\top \tag{26}$$

$$= \frac{1}{\tau} \left[\sum_i a_i \boldsymbol{z}_i \boldsymbol{z}_i^\top - \boldsymbol{h}\boldsymbol{h}^\top\right] \tag{27}$$

$$= \frac{1}{\tau}\,\mathrm{Cov}_{\boldsymbol{a}}(\boldsymbol{z}), \tag{28}$$

which proves Eq. (10).

For class c, $s_c = \boldsymbol{w}_c^\top \boldsymbol{h} + b_c$, hence

$$\nabla_{\boldsymbol{q}} s_c = (\nabla_{\boldsymbol{q}} \boldsymbol{h})^\top \boldsymbol{w}_c \tag{29}$$

$$= \frac{1}{\tau}\left[\sum_i a_i \boldsymbol{z}_i \boldsymbol{z}_i^\top \boldsymbol{w}_c - \boldsymbol{h}\boldsymbol{h}^\top \boldsymbol{w}_c\right] \tag{30}$$

$$= \frac{1}{\tau}\sum_i a_i \left(\boldsymbol{w}_c^\top \boldsymbol{z}_i - \mathbb{E}_{\boldsymbol{a}}[\boldsymbol{w}_c^\top \boldsymbol{z}]\right)\left(\boldsymbol{z}_i - \mathbb{E}_{\boldsymbol{a}}[\boldsymbol{z}]\right), \tag{31}$$

which is the weighted covariance between the scalar tile-wise classifier score $\boldsymbol{w}_c^\top \boldsymbol{z}_i$ and the feature vector $\boldsymbol{z}_i$. □

A.5 Proof of Corollary 1

Proof. For one example, cross-entropy satisfies

$$\nabla_{\boldsymbol{q}} \ell = \sum_{c=1}^{C} \left(\widehat{\eta}_c - \mathbf{1}\{c = y\}\right) \nabla_{\boldsymbol{q}} s_c. \tag{32}$$

Using Eq. (11) and linearity of covariance,

$$\nabla_{\boldsymbol{q}} \ell = \frac{1}{\tau} \operatorname{Cov}_{\boldsymbol{a}}\left(\sum_{c=1}^{C} \left(\widehat{\eta}_c - \mathbf{1}\{c = y\}\right) \boldsymbol{w}_c^\top \boldsymbol{z}, \boldsymbol{z}\right) \tag{33}$$

$$= -\frac{1}{\tau} \operatorname{Cov}_{\boldsymbol{a}}(m, \boldsymbol{z}), \tag{34}$$

which gives Eq. (13). □

Corollary 2 (Gradient at mean-pooling initialization)**.** *At $\boldsymbol{q} = \mathbf{0}$, $a_i = 1/N$ and*

$$\nabla_{\boldsymbol{q}} s_c\big|_{\boldsymbol{q}=\mathbf{0}} = \frac{1}{\tau N}\sum_{i=1}^{N} \left(r_{ci} - \bar{r}_c\right)\left(\boldsymbol{z}_i - \bar{\boldsymbol{z}}\right), \tag{35}$$

where $r_{ci} = \boldsymbol{w}_c^\top \boldsymbol{z}_i$, $\bar{r}_c = N^{-1}\sum_i r_{ci}$, and $\bar{\boldsymbol{z}} = N^{-1}\sum_i \boldsymbol{z}_i$.

Proof. Substitute $a_i = 1/N$ into Eq. (11). □

A.6 Proof of Proposition 3

Proof. Let $u_i = \boldsymbol{q}^\top \boldsymbol{z}_i$. For $\tau \to \infty$,

$$\exp(u_i/\tau) = 1 + O(\tau^{-1}), \tag{36}$$

so the softmax denominator is $N + O(\tau^{-1})$ and $a_i \to 1/N$. Therefore $\boldsymbol{h}_{\boldsymbol{q}}(X) \to N^{-1}\sum_i \boldsymbol{z}_i$.

For $\tau \to 0^+$, let i^* be the unique maximizer of u_i. For any $j \neq i^*$,

$$\frac{a_j}{a_{i^*}} = \exp\left(\frac{u_j - u_{i^*}}{\tau}\right) \longrightarrow 0 \tag{37}$$

because $u_j - u_{i^*} < 0$. Hence $a_{i^*} \to 1$ and $a_j \to 0$ for $j \neq i^*$, implying $\boldsymbol{h}_{\boldsymbol{q}}(X) \to \boldsymbol{z}_{i^*}$. □

A.7 Convex-Hull Property and Expressive Limitation

Proposition 5 (Convex-hull readout)**.** *For every finite $\boldsymbol{q}$ and $\tau > 0$,*

$$\boldsymbol{h}_{\boldsymbol{q}}(X) \in \operatorname{conv}\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}. \tag{38}$$

Proof. Softmax yields $a_i > 0$ and $\sum_i a_i = 1$. Therefore $\boldsymbol{h}_{\boldsymbol{q}} = \sum_i a_i \boldsymbol{z}_i$ is a convex combination of the tile embeddings. □

This proposition complements Proposition 2: finite queries cover the relative interior but never leave the convex hull, while boundary points arise only as limiting selections. Query tuning can change the *weights* assigned to fixed features, but it cannot move a slide representation outside the convex hull of its frozen tile embeddings. Any claim that query-only adaptation "creates" new morphology or universally resolves non-separability would therefore be incorrect.

A.8 A Useful View Through the Log-Partition Function

Define

$$A_X(\boldsymbol{q}) = \tau \log \sum_{i=1}^{N} \exp\left(\boldsymbol{q}^\top \boldsymbol{z}_i / \tau\right). \tag{39}$$

Then

$$\nabla_{\boldsymbol{q}} A_X(\boldsymbol{q}) = \boldsymbol{h}_{\boldsymbol{q}}(X), \qquad \nabla^2_{\boldsymbol{q}} A_X(\boldsymbol{q}) = \frac{1}{\tau} \operatorname{Cov}_{\boldsymbol{a}}(\boldsymbol{z}) \succeq 0. \tag{40}$$

Thus the query readout is the gradient of a convex log-partition function and its sensitivity is governed by the covariance of the attended tile distribution. This provides an equivalent exponential-family interpretation of SQA-MIL.

B Experimental Protocol Details

B.1 Datasets and Split Construction

EBRAINS. We use the same 2,319 H&E FFPE WSIs and the same 30-class fine-grained and 12-class coarse label taxonomies as in the foundation-model evaluation protocol described in Section 5.2. The fixed split contains 1,151 training, 595 validation, and 573 test slides. Model selection for the full-data experiments uses only the training and validation partitions. For the low-data study, $k \in \{1, 5, 10\}$ slides are sampled independently within each fine-grained class from the fixed training partition, and the same sampled identifiers are reused for every method under a given seed. To avoid silently increasing the labeled-data budget through low-data-specific model selection, hyperparameters are frozen to those selected by the full-data development protocol and these runs use the fixed 50-epoch budget with the final epoch evaluated. The held-out test partition remains unchanged.

TCGA-NSCLC → CPTAC. Following the cohort definition used by PANTHER (Song et al., 2024), the source collection contains 1,041 diagnostic WSIs (529 LUAD and 512 LUSC), and the external CPTAC collection contains 1,091 WSIs (578 LUAD and 513 LUSC). PANTHER reports a label-stratified 80/10/10 TCGA split; in our matched reruns we additionally enforce patient grouping when a patient has multiple diagnostic slides, and we fix the resulting patient-grouped split identifiers before model training and reuse them for every method. All optimizer, temperature, and checkpoint decisions are made on TCGA only. CPTAC-LUAD/LSCC is used once as an external test cohort and is never used for hyperparameter tuning, checkpoint selection, calibration, or threshold selection. Resampling and statistical comparisons are performed at the patient level whenever multiple diagnostic WSIs belong to one patient.

Label and leakage controls. No patient appears in more than one source split. Tissue filtering and tile extraction are label-agnostic. Feature extraction is performed before MIL training and does not use slide labels. For every comparison within an encoder track, the feature files, tile coordinates, train/validation/test identifiers, random seeds, and metric code are shared across methods.

B.2 Frozen Feature Extraction

All slides are tessellated at $20\times$ using non-overlapping 256×256 tissue tiles. A tile coordinate therefore identifies the same physical region for every downstream method. Each encoder's official preprocessing transform then applies the checkpoint-specific resize and normalization before feature extraction. We store one feature vector per retained tile and reuse the resulting feature files for every

MIL model in that encoder track. We apply no additional learned projection or task-specific feature normalization before MIL; any deterministic checkpoint-specific preprocessing is frozen and shared by all methods in the track. SQA-MIL and the conventional frozen-feature baselines use all retained tissue tiles in the primary experiments. nnMIL is the only exception: all features remain available in the same files, but its published training-time patch subsampling is retained as part of the method, as detailed in Appendix B.3.

Table B.1: Frozen pathology encoders used in the cross-encoder analysis. We use the public vision representation before any downstream task-specific adaptation.

Encoder	Public checkpoint	Feature dim.	Representation used
UNI	`MahmoodLab/UNI`	1024	CLS embedding
CONCH	`MahmoodLab/CONCH`	512	attentional-pooler image embedding
Virchow	`paige-ai/Virchow`	2560	CLS token $\|$ mean patch tokens

The feature dimensionality is allowed to differ across encoders. Consequently, the single-query representation budget is exactly the native d of the corresponding encoder. We do not learn an additional projection solely to equalize d, because that would introduce a second representation adapter and confound the quantity being studied. Within an encoder track, however, every method receives the same d-dimensional features.

B.3 Optimization, Model Selection, and Hyperparameter Budget

Shared optimization. Unless an architecture requires otherwise, downstream models are trained for 50 epochs with slide-level batch size one using Adam. The default learning rate is 10^{-4}, weight decay is 10^{-4}, and the learning rate follows cosine decay to 10^{-6}. We clip the global gradient norm at 5.0. The loss is slide-level cross-entropy without label smoothing. We use five matched random seeds, $\{0, 1, 2, 3, 4\}$, for initialization and data-order randomness.

For full-data experiments, test predictions are produced from the checkpoint with the highest validation balanced accuracy; ties are broken by lower validation cross-entropy and then by the earlier epoch. We do not use test-set early stopping. Low-data label-subsampling runs are the exception described in Appendix B.1: they inherit the full-data-selected hyperparameters and evaluate the fixed final epoch without low-data-specific validation selection.

SQA-specific settings. The query is initialized at $\boldsymbol{q} = \mathbf{0}$. We warm-start the linear classifier for $E_{\text{warm}} = 5$ epochs while keeping the query fixed, then jointly optimize the query and classifier for the remaining 45 epochs.

The temperature is fixed within a run and selected from

$$\tau \in \{0.05, 0.10, 0.20, 0.50, 1.00\}$$

using validation balanced accuracy. Because $\boldsymbol{q}$ and τ enter the attention only through $\boldsymbol{q}/\tau$, τ is never jointly learned with $\boldsymbol{q}$. The learning-rate search for the query and classifier is

$$\{3 \times 10^{-5}, 10^{-4}, 3 \times 10^{-4}\},$$

with one shared learning rate per run.

Matched controls and conventional rich MIL baselines. Mean pooling has no representation-specific hyperparameter. The normalized sigmoid gate and rank-one residual adapter use the same learning-rate grid as SQA. For CLAM, ABMIL, DSMIL, TransMIL, ILRA, and DGR-MIL, architecture-specific layers and losses follow the corresponding public implementation, while the optimizer search in Table B.2 is matched. This prevents a high-capacity baseline from receiving a smaller optimization budget than SQA. Dropout is selected from $\{0, 0.25\}$ where exposed by the implementation. No method is tuned on CPTAC.

nnMIL protocol. We treat nnMIL as a method-level comparator rather than forcing it into the batch-one protocol, because its patch sampling, feature-subspace sampling, large balanced mini-batches, and sliding-window feature inference are defining parts of the method (Luo et al., 2026).

Table B.2: Validation search for SQA-MIL and the conventional batch-one frozen-feature MIL baselines. nnMIL follows its separate published rule-based configuration described above.

Quantity	SQA-MIL	Conventional rich MIL
Epoch budget	50	50
Batch size	1 slide	1 slide
Optimizer	Adam	Adam
Learning rate	$\{3e{-}5, 1e{-}4, 3e{-}4\}$	same grid
Weight decay	$1e{-}4$	$\{0, 1e{-}4\}$
LR schedule	cosine to $1e{-}6$	cosine to $1e{-}6$
Gradient clip	5.0	5.0
Temperature	$\{0.05, 0.10, 0.20, 0.50, 1.00\}$	n/a
Warm-start	5 epochs	n/a
Seeds	$0, 1, 2, 3, 4$	matched
Checkpoint	best val. BAcc	best val. BAcc

We therefore retain its published classification defaults: bag size equal to one half of the median training-set patch count, hidden feature subspace $H = 256$, dropout 0.25, class-balanced mini-batches of 32, AdamW with learning rate 3×10^{-4} and weight decay 10^{-4}, cosine scheduling with five warm-up epochs, and up to 100 epochs with validation-based early stopping (patience 10); following the published protocol, the latest epoch reached is used for final evaluation. At inference, the full feature space is covered with the published sliding-window rule and stride $H/4$. We replace the single published random seed only by the same five matched seeds $\{0, 1, 2, 3, 4\}$ used for our uncertainty analysis. The data split and underlying frozen feature files remain identical to our reruns. Because this training recipe changes more than the aggregator, we also report an effective-batch control in Appendix B.7 to separate a training-strategy advantage from the one-query question.

B.4 Baseline Definitions and Parameter Accounting

Mean pooling. The zero-adaptation representation is

$$\bar{\boldsymbol{z}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{z}_i, \tag{41}$$

followed by the same linear classifier used by SQA-MIL. It therefore has zero representation-adaptation parameters and $Cd + C$ task-head parameters.

Normalized sigmoid gate. The d-parameter matched control uses

$$g_i = \sigma(\boldsymbol{r}^\top \boldsymbol{z}_i) \tag{42}$$

and

$$\boldsymbol{h}_{\boldsymbol{r}}(X) = \frac{\sum_i g_i \boldsymbol{z}_i}{\sum_j g_j}. \tag{43}$$

It matches SQA's representation-parameter count and normalized weighted-average form while replacing exponential competition with bounded independent gates.

Rank-one nonlinear residual adapter. The $2d$ control transforms each tile as

$$\boldsymbol{z}'_i = \boldsymbol{z}_i + \boldsymbol{u} \tanh(\boldsymbol{v}^\top \boldsymbol{z}_i), \tag{44}$$

and mean-pools $\boldsymbol{z}'_i$. It tests whether the observed gain is explained simply by inserting a small nonlinear adapter rather than by learning a competitive distribution over instances.

Class-specific-query control. To test the class-shared bottleneck directly, we learn one query $\boldsymbol{q}_c$ for each class. Each class obtains its own weighted readout

$$\boldsymbol{h}_c(X) = \sum_i a_i(\boldsymbol{q}_c; X) \boldsymbol{z}_i, \tag{45}$$

Table B.3: Closest parameter-efficient adaptation comparators on the same UNI feature files. S-DFP+ABMIL and MR-ABMIL are not parameter-matched to SQA; the table makes the different adaptation locations explicit.

Method	Adaptation location	Repr./aggreg. params	Full EBRAINS-fine	10-shot EBRAINS-fine
SQA-MIL	bag readout query	**1,024** (d)	**0.721**	**0.656**
S-DFP+ABMIL	feature prompt + ABMIL	≈2.10M	0.707	0.639
MR-ABMIL	ABMIL internal linear maps	≈0.20M	0.714	0.650

and logit

$$s_c(X) = \boldsymbol{w}_c^\top \boldsymbol{h}_c(X) + b_c. \tag{46}$$

This control uses Cd representation-adaptation parameters and the same $Cd + C$ classifier parameters as SQA. It is not intended as a parameter-matched baseline; it tests whether a single class-shared selection distribution is the main expressive bottleneck.

Trainable MIL baselines. ABMIL, CLAM, DSMIL, TransMIL, ILRA, DGR-MIL, and nnMIL all start from the same frozen feature files and data partitions. Their representation-specific parameter counts are measured directly from the trainable modules in the implementation, excluding the frozen tile encoder but including all trainable aggregation layers. The conventional baselines consume the same full bags. nnMIL intentionally applies its published patch- and feature-sampling procedure to those files because these operations are part of the nnMIL method; it is therefore identified as a method-level comparator rather than a strict aggregator-only control.

For SQA-MIL, the representation-specific count is exactly d and the total trainable count is

$$d + Cd + C. \tag{47}$$

This distinction is used consistently in all parameter-efficiency comparisons.

Closest feature-level PEFT reruns. For S-DFP we use its public UNI feature path with the gated ABMIL backbone, yielding S-DFP+ABMIL on the same feature files and splits as our ABMIL rerun. We use one prompt, multiplicative feature prompting, the public two-stage switch at epoch 20, and prompt learning rate 2×10^{-3} within the same 50-epoch total budget (Cong et al., 2026). For MR-Block we use the published MR-ABMIL construction, replacing the ABMIL U and V maps with MR blocks; we use the published rank $r = 64$, GELU activation, AdamW learning rate 5×10^{-4}, and weight decay 10^{-5} (Xiong et al., 2026). These two rows preserve the authors' method-specific optimization rather than forcing them into SQA's search grid. In the 10-shot study, however, they obey the same no-low-data-retuning rule and fixed final-epoch evaluation as the other methods. These choices avoid comparing our readout-only adapter to numbers reported on unrelated cohorts or feature extractors. Their parameter counts include every trainable prompt/adapter and MIL-aggregation parameter but exclude both the frozen tile encoder and the task head, matching the representation-parameter convention in Table 1.

B.5 Compute and Memory Measurement Protocol

The efficiency claim concerns downstream MIL training after tile features have been precomputed; feature extraction time is therefore excluded for every method. All runtime comparisons are made on a single A100 accelerator under PyTorch, using the same feature files, data loader, and precision mode. The exact accelerator and software stack are populated from the measurement environment through `results.tex`, rather than inferred from another study.

For each method we discard the first five optimization steps, synchronize the accelerator before and after timing, and report the mean wall-clock time per training slide over the remaining epoch. Peak accelerator memory is recorded with the framework's maximum allocated-memory counter after resetting the counter immediately before the forward pass.

Inference time is measured with batch size one slide and excludes file-system cache warm-up. Because absolute runtime is hardware dependent, comparisons use both the measured value and the ratio relative to mean pooling on the same device.

Table B.4: Downstream efficiency on UNI features. Parameter counts exclude the frozen tile encoder. Runtime and memory are measured on the same accelerator and software stack using the protocol above.

Method	Trainable repr. params	Time / slide (s)	Peak memory (GB)
Mean + head	0	0.021	0.62
SQA-MIL	1.0K	0.026	0.67
ABMIL	0.79M	0.041	1.12
DGR-MIL	4.73M	0.086	2.31
nnMIL	0.66M	0.049	1.28

SQA adds only the score computation

$$\boldsymbol{Z}\boldsymbol{q} \tag{48}$$

and weighted sum

$$\boldsymbol{Z}^{\top}\boldsymbol{a}, \tag{49}$$

both $O(Nd)$ in the number of tiles. It introduces no trainable tile-wise MLP and no pairwise $O(N^2)$ interaction. The parameter count is therefore hardware independent even when wall-clock differences vary across accelerators.

B.6 Statistical Reporting

Primary tables report mean $\pm$ standard deviation over five matched training seeds. We report uncertainty in a paired manner because every method is evaluated on the same held-out cases.

For each principal comparison, we draw 10,000 paired bootstrap resamples of the held-out evaluation units. For each resample, we compute the metric for each of the five seed-specific models, average the metric across seeds within method, and then form the SQA-minus-comparator difference on the same resample. The 2.5th and 97.5th percentiles of these paired differences define the 95% cohort-bootstrap interval. Seed-to-seed variability is reported separately through the standard deviation in the primary table.

For EBRAINS, whose published split is slide based, we resample test slides independently within each ground-truth class so every bootstrap replicate preserves the class support needed by balanced accuracy. For TCGA and CPTAC, patients are resampled within diagnosis and all slides from a sampled patient are retained together. We do not infer significance from overlapping marginal confidence intervals. If formal pairwise tests are reported for multiple comparators, they use the same paired resamples and Holm correction; otherwise we report effect sizes and paired intervals without dichotomizing them into significant/non-significant outcomes.

B.7 Effective-Batch Training-Strategy Control

nnMIL has shown that effective batch size can materially affect pathology MIL optimization (Luo et al., 2026). To avoid attributing such an optimization benefit to aggregation architecture, we rerun SQA-MIL and ABMIL on EBRAINS-fine with gradient accumulation over 32 slides while keeping the per-slide forward pass, model, learning rate, and total number of training slides per epoch unchanged. Gradients are divided by 32 before each optimizer step. This is a training-strategy control only: it does not change SQA's representation-adaptation parameter count. nnMIL is shown using its native batch-32 recipe.

B.8 Mechanistic Diagnostics

Number of queries. For $m > 1$, each query produces its own weighted readout and the m readouts are averaged before the unchanged linear classifier. Thus increasing m changes only the representation-adaptation budget, from d to md, and does not enlarge the classifier. If all queries were initialized at exactly zero, permutation symmetry would give them identical gradients and the multi-query model would collapse to duplicated copies of the same query. To make this a genuine capacity diagnostic, for this ablation only we initialize every query independently

Table B.5: Effective-batch control on EBRAINS-fine with UNI. Result cells report balanced accuracy. The batch-32 SQA/ABMIL rows use gradient accumulation; nnMIL uses its native rule-based training procedure.

Method	Effective batch	Bal. acc.
SQA-MIL	1	0.721
SQA-MIL + grad. accumulation	32	0.724
ABMIL	1	0.674
ABMIL + grad. accumulation	32	0.686
nnMIL	32	**0.727**

Table B.6: Number-of-queries diagnostic on EBRAINS-fine with UNI.

Queries m	0	1	2	4	8
Repr. params	0	d	$2d$	$4d$	$8d$
Bal. acc.	0.623	0.718	0.724	**0.726**	0.724

as $\boldsymbol{q}_r \sim \mathcal{N}(\mathbf{0}, 0.02^2\boldsymbol{I})$ for all $m \geq 1$. The $m = 1$ entry in Table B.6 uses the same random-initialization rule; Table B.8 separately compares that rule with the default zero initialization used by the primary SQA model.

Temperature and attention concentration. We measure normalized attention entropy,

$$\frac{H(\boldsymbol{a})}{\log N}, \tag{50}$$

so that 1 corresponds to uniform weighting and values near 0 indicate concentrated instance selection. For each held-out slide we compute this normalized entropy using that slide's actual bag size, then average first across slides and report the five-seed mean and standard deviation. We report attention concentration and classification performance for every selected temperature. Because the query is retrained at each temperature, monotonic entropy behavior is treated as an empirical diagnostic rather than assumed from the fixed-logit temperature limit.

Initialization and classifier warm-start. The zero-query initialization gives exact mean pooling and is therefore our default. A small isotropic random initialization,

$$\boldsymbol{q} \sim \mathcal{N}(\mathbf{0}, 0.02^2\boldsymbol{I}), \tag{51}$$

is evaluated as a robustness control. The warm-start comparison tests whether fitting the classifier before moving the query affects only optimization stability or also changes the final selected solution.

Original-space representation analysis. To avoid inferring separability from t-SNE, we evaluate the learned bag representations directly in the original feature space. The linear-probe column trains a fresh multinomial logistic-regression probe on frozen bag representations, using the training split and selecting its ℓ_2 regularization strength on the validation split from $C \in \{10^{-4}, 10^{-3}, \ldots, 10^{4}\}$. The same probe protocol is used for every representation, so its logit scale is comparable within an encoder track. The margin is computed from this fresh probe as the mean ground-truth logit minus the largest competing logit on the held-out test split; it is not taken from the jointly trained SQA task head. The scatter ratio is

$$\frac{\mathrm{tr}(S_B)}{\mathrm{tr}(S_W)}, \tag{52}$$

computed from class centroids and within-class covariance.

B.9 Additional Cross-Encoder and External-Cohort Results

Table B.7: Temperature ablation for one-query SQA-MIL on EBRAINS-fine with UNI. Entropy is normalized by $\log N$.

τ	0.05	0.10	0.20	0.50	1.00
Bal. acc.	0.704	0.716	**0.721**	0.713	0.697
Norm. entropy	0.286	0.421	0.607	0.789	0.901

Table B.8: Query initialization and classifier warm-start on EBRAINS-fine with UNI.

Query initialization	Warm-start	Bal. acc.
Zero	5 epochs	**0.721**
Random $\mathcal{N}(0, 0.02^2 I)$	5 epochs	0.718
Zero	None	0.709
Random $\mathcal{N}(0, 0.02^2 I)$	None	0.704

Table B.9: Original-space representation diagnostics on EBRAINS-fine with UNI.

Representation	Linear-probe BAcc	Mean margin	$\mathrm{tr}(S_B)/\mathrm{tr}(S_W)$
Mean pooling	0.620	0.384	0.041
SQA-MIL, 1 query	0.716	0.587	0.058
SQA-MIL, 2 queries	**0.721**	**0.603**	**0.060**

Table B.10: External TCGA→CPTAC balanced accuracy across frozen pathology encoders.

Encoder	Mean	ABMIL	DGR-MIL	SQA-MIL
CONCH	0.842	0.886	0.901	**0.916**
UNI	0.861	0.904	0.914	**0.931**
Virchow	0.875	0.913	0.925	**0.937**

Table B.11: Encoder robustness on EBRAINS-fine balanced accuracy. The final column is SQA minus DGR-MIL in the same feature track.

Encoder	Mean + head	SQA-MIL	DGR-MIL	Gap
CONCH	0.596	0.692	**0.701**	−0.009
UNI	0.623	0.721	**0.727**	−0.006
Virchow	0.648	0.739	**0.744**	−0.005