QG-MIL: A Gated Transformer Aggregator for Domain-Agnostic Multiple Instance Learning in Medical Imaging

Luca Zedda[1*†], Davide Antonio Mura[1*], Cecilia Di Ruberto[1], Maurizio Atzori[1], Muhammed Furkan Dasdelen[2], Carsten Marr[2], and Andrea Loddo[1]

[1] Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy
[2] Institute of AI for Health, Helmholtz Munich, Neuherberg, Germany

[*]Contributed equally to this work and share first authorship.
[†]Corresponding author: luca.zedda@unica.it

Abstract. Attention-based Multiple Instance Learning aggregators in medical imaging are prone to attention concentration, producing overconfident and unstable predictions. We introduce QG-MIL, a gated transformer aggregator that addresses this through four synergistic architectural components: RMSNorm-based pre-normalization, per-head QK normalization, fine-grained attention output gating, and SwiGLU-style feed-forward modules. Together, these design choices stabilize training and distribute attention more uniformly across instances without auxiliary losses, masking, or multi-stage regularization. We evaluate QG-MIL across six benchmarks spanning whole-slide pathology and cell-level hematology, covering two fundamentally different MIL scales. The best-performing QG-MIL variants outperform leading baselines on all six benchmarks, with an average improvement of +6.1 mean macro F1 points. Attention overlays and attention mass analysis confirm more distributed instance weighting. Ablation studies show that while individual components can match the full model on specific datasets, the QG-MIL design provides the most consistent cross-domain performance and tightest variance when compared to selected baselines. We release a configurable implementation to support reproducibility at:
https://github.com/unica-visual-intelligence-lab/QG-MIL

Keywords: Multiple Instance Learning · Weakly Supervised Classification· Gated Transformer· Digital Pathology· Hematology

1 Introduction

Multiple Instance Learning (MIL) has become a core paradigm for computational pathology and medical imaging tasks where slide- or patient-level labels are available but instance-level annotations are costly. Attention-based MIL methods learn to aggregate instance features by assigning weights that both drive predictions and serve as a proxy for instance importance [9], making them

attractive for clinical pipelines that require some degree of interpretability. A well-documented failure mode of these methods is attention concentration: the learned weights collapse onto a handful of instances, creating attention sinks that yield overconfident predictions and degrade generalization across cohorts and imaging sources [20]. Existing remedies operate at the training level (attention masking [20], self-supervised pretraining [11], or teacher-student distillation [14, 17]). While effective, they often introduce additional stages or auxiliary losses that complicate the overall pipeline. Rather than regularizing attention after the fact, we redesign the aggregation module itself so that concentrated attention is structurally discouraged. Our initial observation is that attention concentration in MIL closely resembles the information collapse phenomenon recently investigated in large-scale language models [13], where gating and normalization inside the attention block have proven effective countermeasures. Therefore, we translate these ideas into a compact MIL aggregator QG-MIL (Qwen Gated Multiple Instance Learning) that slots into any existing pipeline as a drop-in replacement in standard MIL setups, with no extra training stages or loss terms. A key point of this work is domain agnosticism. MIL problems in medical imaging range from gigapixel whole-slide images with thousands of patches to blood-smear analysis with hundreds of single-cell instances. An aggregation module that works well in one regime but not the other is of limited practical value. Therefore, we evaluate QG-MIL on both pathology and hematology benchmarks, using the same architecture and hyperparameters throughout, to stress-test generalization across fundamentally different bag sizes and imaging modalities. Our main contributions are:

- We propose QG-MIL, a gated transformer aggregation module for MIL that mitigates attention concentration through architectural design alone, without auxiliary losses or multi-stage training, along with implementation code.
- We evaluate QG-MIL on six benchmarks across pathology and hematology, showing consistent improvements in predictive performance, attention distribution, and localization quality with the QG-MIL model and its ablations.
- We provide ablation studies isolating each design choice and show the impact of the cohort size and model depth on ablation performances.

2 Methodology

In MIL, each sample is a bag $\mathcal{B} = \{x_i\}_{i=1}^{N}$ with instance features $x_i \in \mathbb{R}^d$ and a single bag label y. QG-MIL maps instances in a bag to a shared latent space, processes them with L stacked gated transformer blocks, and aggregates them through gated attention pooling to obtain a bag representation for classification. We denote W_p as the input projection; Q, K, and V as the query, key, and value projections; A as the attention weights; Z as the attention output before gating; O as the gated attention output; and W_o as the output projection.

Instances are first projected to dimension D as $h_i = \mathrm{Dropout}(\mathrm{RMSNorm}(W_p x_i))$, yielding $H = [h_1; \ldots; h_N] \in \mathbb{R}^{N \times D}$. For each block, linear projections produce $\mathcal{Q}, \mathcal{K}, \mathcal{V}$, which are reshaped into multi-head form with M attention heads and

per-head dimension d_h, such that $D = Md_h$. Queries and keys are normalized per head for stability ($\widehat{Q}$, $\widehat{K}$), and scaled dot-product attention is computed as

$$S = \frac{\widehat{Q}\,\widehat{K}^{\top}}{\sqrt{d_h}}, \qquad A = \mathrm{softmax}(S). \tag{1}$$

The attention output of head h for token i is $Z_{i,h,:} = A_{i,h}V_{:,h,:} \in \mathbb{R}^{d_h}$. Gating is applied to Z before the output projection. Crucially, the gate activations are computed via dedicated learnable linear projections parameterized by weights $W_{g,h}$. In the headwise formulation, the gate is computed via a projection to a scalar:

$$z_{i,h} = W_{g,h}Z_{i,h,:}, \qquad g_{i,h} = \sigma(z_{i,h}) \tag{2}$$

which modulates all features uniformly: $O_{i,h,:} = g_{i,h}Z_{i,h,:}$. In the elementwise formulation, the projection maintains the feature dimension d_h, modulating each feature independently: $O_{i,h,:} = g_{i,h,:} \odot Z_{i,h,:}$.

Concatenating heads and applying the output matrix yields the block attention output:

$$\mathrm{Attn}(H) = W_o\big(\mathrm{concat}_h O_{:,h,:}\big). \tag{3}$$

Each block follows a pre-norm residual layout with a SwiGLU feed-forward network $\mathrm{FFN}(x) = W_d\big(\phi(W_g x)\odot W_u x\big)$ and updates $x' = x + \mathrm{Dropout}(\mathrm{Attn}(\mathrm{Norm}(x)))$ and $x'' = x' + \mathrm{Dropout}(\mathrm{FFN}(\mathrm{Norm}(x')))$. Stacking L such blocks yields refined instance embeddings.

Finally, gated attention pooling aggregates the final processed instances h_i into a bag representation,

$$a_i = w^{\top}\big(\tanh(W_v h_i)\odot\sigma(W_u h_i)\big), \qquad \alpha_i = \frac{\exp(a_i)}{\sum_j \exp(a_j)}, \qquad z = \sum_i \alpha_i h_i, \tag{4}$$

and z is fed to a classification head to predict the bag label.

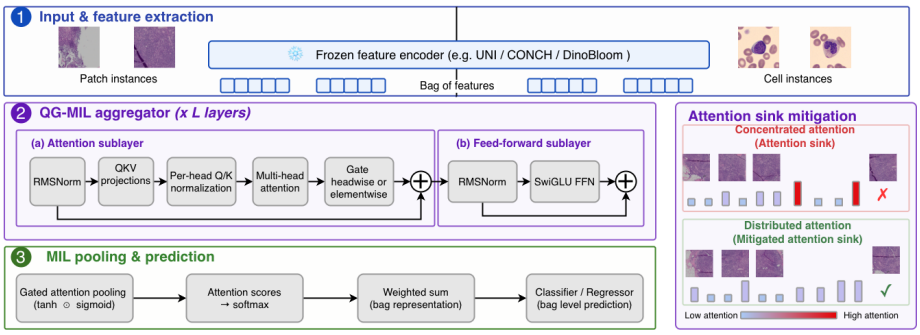

Fig. 1. Overview of the proposed QG-MIL architecture.

3 Experiments and Results

Evaluation Data: For histopathology, we utilize two diagnostic whole-slide datasets: MSK [1](Breast): 130 WSIs from 78 patients, 36 with metastatic carcinoma, and LungHist700 [5](Lung): 691 images from 45 patients across normal and carcinoma classes. Additionally, we include a prognostic benchmark: Prostate Cancer [2](Prostate). This dataset comprises 587 biopsies from 213 initially benign patients, focusing on predicting future cancer development, cancer-free ≥ 8 years versus prostate cancer within 30 months. To assess generalizability, the Breast and Lung datasets are encoded with foundation models: UNI2-h [3], Prov-Gigapath [18], and CONCHv1.5 [12]. The challenging prognostic prostate dataset is additionally processed with UNIv1 [3] and ResNet50 [6] to evaluate QG-MIL across both foundation and conventional CNN extractors. For cell-level hematology, we evaluate on three benchmark datasets: *AML-Hehr* [7] 129 patients distributed in 5 classes, *APL-AML* [16] 106 patients grouped in 2 classes, and *cAItomorph* [4] 2043 patients across 8 classes. These are processed using DinoBloom [10], a white blood cell specialized foundation model, across two encoder sizes, specifically Small and Large. All patch and cell encoders are frozen and used as static feature extractors.

Experimental protocol. To ensure robust evaluation across varying cohort sizes, we allocate a patient-stratified 20% hold-out test set for final assessment. Within the remaining 80% of the patient cohort, we employ a 5-fold cross-validation strategy to train five independent models. During inference, the predictions from these fold-specific models are ensembled on the fixed test set using mean probability aggregation. Our data splitting and training regimen are explicitly designed to promote fair evaluation by adopting a shared scenario that facilitates direct comparisons with existing literature. By utilizing these common benchmark settings whenever applicable, we strictly adhere to the experimental parameters and patient splits established by previous works to ensure the reproducibility of our findings. Demonstrating this commitment to a shared evaluation framework, we utilize the original data splits defined by [4] for all hematology datasets. Finally, to quantify model stability, we report the standard deviation of the test-set performance across the five-fold-specific models.

Given the pronounced class imbalance typical of medical datasets, we select macro F1 as the primary evaluation metric. All models are trained using CrossEntropy loss for a maximum of 150 epochs, with early stopping applied after 10 epochs without improvement in validation loss, dropout of 0.25, and embedding size D of 512. Model selection for each fold is based exclusively on the lowest validation loss. To maintain architectural consistency and isolate the contribution of the proposed design, all QG-MIL variants are implemented with two layers, matching the depth of baseline MIL aggregators, including ILRA, RRT, Transformer, and TransMIL, following the standard implementation of [15]. While further parameter tuning could potentially improve absolute performance, exhaustive optimization is beyond the scope of this study. Our goal is instead to assess architectural behavior under controlled and comparable conditions.

Table 1. Summary classification performance on Breast and Lung pathology benchmarks. QG-MIL consistently improves F1. Best results per column are bolded; second-best are underlined.

Model	Breast				Lung			
	UNI2-h	Prov-Gigapath	CONCHv1.5	Mean	UNI2-h	Prov-Gigapath	CONCHv1.5	Mean
ABMIL	81.21 ± 3.3	81.01 ± 5.9	91.30 ± 5.4	84.51	90.80 ± 3.1	85.56 ± 4.0	86.72 ± 2.0	87.69
CLAM	81.21 ± 3.3	79.72 ± 4.1	91.30 ± 5.4	84.08	89.94 ± 3.2	84.12 ± 3.87	85.03 ± 2.6	86.36
DSMIL	82.50 ± 4.9	81.13 ± 18.2	93.73 ± 2.6	85.79	91.11 ± 3.1	84.45 ± 3.2	84.48 ± 3.9	86.68
ILRA	75.99 ± 10.0	81.16 ± 9.4	87.72 ± 5.1	81.62	88.74 ± 3.7	80.70 ± 3.5	86.34 ± 1.9	85.26
RRT	82.50 ± 4.9	81.01 ± 5.9	94.87 ± 0.0	86.13	92.17 ± 2.1	86.63 ± 2.70	84.68 ± 4.71	87.83
Transformer	77.95 ± 7.2	85.09 ± 6.2	91.30 ± 5.4	84.78	90.63 ± 2.6	86.79 ± 2.9	85.02 ± 4.2	87.48
TransMIL	81.21 ± 3.3	79.24 ± 8.7	90.01 ± 6.7	83.49	90.58 ± 3.2	86.54 ± 4.5	83.83 ± 3.7	86.98
WIKG	82.71 ± 0.0	**91.51 ± 5.3**	91.30 ± 5.4	88.51	91.95 ± 4.3	84.75 ± 3.0	87.91 ± 3.0	88.20
MeanMIL	53.66 ± 13.0	53.82 ± 12.5	41.96 ± 0.6	49.81	86.28 ± 15.0	73.77 ± 13.9	85.25 ± 9.4	81.77
QG-MIL	**85.29 ± 3.5**	87.72 ± 5.1	91.30 ± 5.4	88.10	93.00 ± 3.5	87.39 ± 1.9	87.45 ± 3.2	89.28
QG-MIL Deep	**85.29 ± 3.5**	87.52 ± 7.3	**98.97 ± 2.3**	**90.59**	92.65 ± 1.8	86.10 ± 2.5	85.52 ± 3.1	88.09
QG-MIL Elementwise	82.30 ± 7.0	90.16 ± 5.1	91.30 ± 5.4	87.92	92.70 ± 2.0	86.14 ± 3.4	87.98 ± 2.7	88.94
QG-MIL Layernorm	84.00 ± 2.9	91.45 ± 3.1	95.78 ± 4.5	90.41	92.67 ± 3.3	87.52 ± 2.6	85.89 ± 4.2	88.69
QG-MIL Light	81.21 ± 3.3	89.02 ± 4.3	95.90 ± 2.3	88.71	90.30 ± 8.7	86.29 ± 6.1	86.27 ± 1.8	87.62
QG-MIL no Gate	**85.29 ± 3.5**	86.23 ± 7.5	95.90 ± 2.3	89.14	**93.34 ± 3.7**	**87.78 ± 4.5**	**88.30 ± 4.9**	**89.81**
QG-MIL no QKnorm	84.00 ± 2.9	91.45 ± 3.1	95.78 ± 4.5	90.41	92.09 ± 1.9	87.54 ± 3.8	85.89 ± 4.2	88.51
Mean Baselines	77.66	79.30	85.94	80.97	90.24	83.70	85.47	86.47
Mean QG-MIL	**83.91**	**89.08**	**94.99**	**89.33**	**92.39**	**86.97**	**86.76**	**88.71**

Baselines and ablations. Based on this architecture, we evaluate six ablation variants: *QG-MIL Elementwise*, which modifies gating granularity; *QG-MIL noGate*, which removes attention-output gating; *QG-MIL noQKnorm*, which disables Q/K normalization; *QG-MIL LayerNorm*, which replaces RMSNorm with LayerNorm; *QG-MIL Light*, which reduces the hidden dimension; and *QG-MIL Deep*, which increases depth. The standard QG-MIL model has 9M parameters, while the Light variant has 2.4M. Figure 1 summarizes the single-stage pipeline from frozen instance features to gated transformer aggregation and bag-level prediction.

General Architectural and Performance Findings. Across both computational pathology and hematology domains, consistent patterns emerge regarding the behavior and efficacy of the proposed QG-MIL framework. First, we observe that the optimal architectural complexity is strictly dictated by cohort size. In low-patient settings, stringent gating mechanisms and Q/K normalization introduce variance that degrades generalization, indicating that smaller datasets achieve optimal performance when regularization constraints are relaxed. Conversely, medium-to-large cohorts remain highly stable during optimization and benefit significantly from the full gating architecture.

Second, QG-MIL exhibits a strong positive correlation between network depth and predictive performance. Assuming sufficient data, deeper aggregation variants, such as QG-MIL Deep, consistently unlock superior macro F1 scores, demonstrating that the architecture scales effectively without suffering from severe optimization bottlenecks. Finally, the refined aggregation mechanisms prove exceptionally adept at capturing subtle morphological features. This capability drives consistent performance improvements over baseline MIL methods, such as standard Transformers and WIKG, and becomes distinctly advantageous in highly constrained, early-stage prognostic scenarios.

Pathology Benchmarks on Diagnosis. The diagnostic benchmarks contrast a low-patient Lung dataset with a medium-to-large Breast dataset. Reflecting our

Table 2. Classification performance on the Prostate Cancer prognostic task. The QG-MIL LayerNorm variant achieves the highest overall average (57.55%), demonstrating superior early prognostic capability compared to baseline aggregators. Best results per column are bolded; second-best are underlined.

Model	Prostate					
	CONCHv1.5	**Prov-Gigapath**	**ResNet50**	**UNIv1**	**UNI2-h**	**Mean**
ABIMIL	56.09 ± 3.0	56.09 ± 4.8	44.16 ± 0.0	59.68 ± 4.9	56.09 ± 2.5	54.42
CLAM	51.80 ± 6.9	56.30 ± 1.4	44.10 ± 0.0	57.20 ± 4.7	57.50 ± 2.3	53.38
DSMIL	**59.10 ± 7.6**	58.00 ± 2.6	49.90 ± 4.7	59.80 ± 3.2	53.90 ± 4.2	56.14
ILRA	52.75 ± 5.0	56.09 ± 4.5	54.39 ± 2.3	59.68 ± 2.2	54.39 ± 3.8	55.46
RRT	56.09 ± 4.3	49.33 ± 5.8	50.82 ± 3.6	54.39 ± 5.3	54.39 ± 2.9	53.00
Transformer	54.39 ± 7.6	**60.91 ± 3.2**	56.09 ± 5.6	59.05 ± 5.1	54.39 ± 2.5	56.97
TransMIL	54.39 ± 7.5	54.39 ± 4.2	56.09 ± 4.4	56.09 ± 4.0	56.09 ± 5.6	55.41
WIKG	54.39 ± 2.9	54.39 ± 3.7	56.09 ± 3.9	54.39 ± 5.3	52.75 ± 6.1	54.40
MeanMIL	44.16 ± 0.0	47.17 ± 6.7	44.16 ± 0.0	50.49 ± 6.6	54.07 ± 6.0	48.01
QG-MIL	56.09 ± 1.5	54.39 ± 3.5	52.75 ± 5.7	56.09 ± 1.6	**59.68 ± 1.9**	55.80
QG-MIL Deep	56.09 ± 1.7	54.39 ± 2.5	49.33 ± 5.8	56.09 ± 5.4	57.84 ± 3.2	54.75
QG-MIL Elementwise	56.09 ± 1.5	57.84 ± 3.6	54.39 ± 2.6	57.84 ± 3.9	56.09 ± 3.1	56.45
QG-MIL Layernorm	57.84 ± 2.2	54.39 ± 4.1	**57.84 ± 3.7**	**61.61 ± 5.2**	56.09 ± 3.0	**57.55**
QG-MIL Light	54.39 ± 3.7	54.39 ± 3.8	51.14 ± 2.2	57.84 ± 2.2	56.09 ± 2.8	54.77
QG-MIL no Gate	47.51 ± 4.8	54.39 ± 2.1	49.33 ± 2.9	56.09 ± 5.1	57.84 ± 2.6	53.03
QG-MIL no QKnorm	56.09 ± 2.3	54.39 ± 1.9	**57.84 ± 3.7**	59.68 ± 3.8	56.09 ± 3.5	56.82
Mean Baselines	53.68	54.74	50.64	56.75	54.84	54.13
Mean QG-MIL	**54.87**	**54.88**	**53.23**	**57.89**	**57.10**	**55.59**

general findings, the Lung benchmark favors relaxed configurations, achieving a peak macro F1 of 93.3%. This establishes a substantial performance margin over previous studies, which reported classification accuracies of 81.6% [19] and 81.0% [5]. In contrast, the larger Breast cohort leverages deeper aggregation effectively. The QG-MIL Deep variant achieves the highest macro F1 of 98.9%. Across all encoders and folds, the QG-MIL variants consistently achieve higher mean macro-F1 scores, with only WIKG showing comparable average baseline performance. Full results are reported in Table 1.

Pathology Benchmarks on Prognosis. The Prostate dataset introduces a highly challenging prognostic task: predicting the future onset of malignancy in patients initially diagnosed as strictly benign. Due to the inherent difficulty of early prediction, overall metrics are markedly lower across all models. While baseline methods achieve an average macro F1 score of 54.1% (peaking at 56.9% for the standard Transformer), the QG-MIL Layernorm variant achieves the highest overall average macro F1 of 57.5%. This highlights the framework's superior ability to isolate the elusive, early-stage morphological signals required to predict malignant transformations prior to standard clinical diagnosis (Table 2).

Hematology Benchmarks on Diagnosis. Beyond solid tissue pathology, QG-MIL demonstrates robust efficacy in hematological tasks, outperforming baseline methods by an average of over 5% in macro F1 score. For the classification of Acute Promyelocytic Leukemia (APL) versus non-APL, the model achieves a strong macro F1 of 69.5% utilizing the deep configuration, directly mirroring the depth-to-regularization trend observed in the Lung dataset. In Acute Myeloid Leukemia (AML) subtyping, QG-MIL reaches a macro F1 of 86.0%, surpassing other specialized MIL methods [8] by 5%. Furthermore, on the cAItomorph dataset, the QG-MIL Deep variant achieves a maximum weighted F1 of 67.9%,

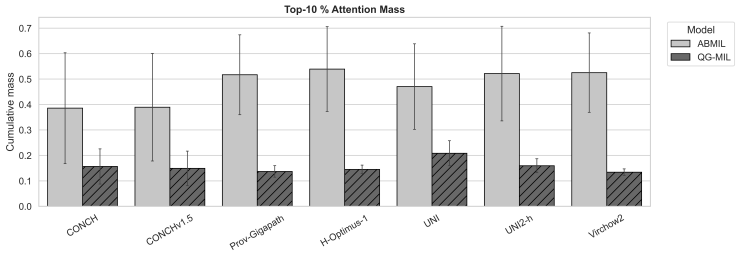

Fig. 2. Mean top-10 attention mass extracted from QG-MIL and ABMIL models on the Breast dataset test set. Lower top-k mass indicates more distributed attention and reduced attention sink. QG-MIL variants show reduced concentration compared to ABMIL, suggesting improved instance coverage. Error bars denote standard deviation across test set images.

representing a 3.7% performance increase over previous studies evaluating up to 8-layer transformer models [4]. Full results are reported in Table 3.

Table 3. Classification performance macro F1 score across cell-level hematology benchmarks. Results demonstrate the domain-agnostic scalability of QG-MIL, as it outperforms established baselines by an average of >5% across varying hematological cell analysis tasks. Best results per column are bolded; second-best are underlined.

Model	AML-Hehr			APL-AML			cAItomorph		
	Bloom-S	Bloom-L	Mean	Bloom-S	Bloom-L	Mean	Bloom-S	Bloom-L	Mean
ABMIL	67.79 ± 6.8	69.81 ± 8.5	68.80	59.01 ± 6.8	39.50 ± 14.1	49.26	57.23 ± 2.7	55.40 ± 3.2	56.32
TransMIL	72.59 ± 8.1	64.86 ± 9.7	68.72	39.50 ± 8.6	39.50 ± 9.4	39.50	58.24 ± 3.9	57.83 ± 1.6	58.04
Transformer	76.97 ± 9.1	81.14 ± 8.0	79.06	64.44 ± 6.8	64.44 ± 3.4	64.44	66.39 ± 2.3	63.45 ± 1.5	64.92
DSMIL	63.11 ± 10.1	73.14 ± 4.0	68.12	41.98 ± 3.3	49.58 ± 9.6	45.78	52.02 ± 3.8	51.76 ± 2.9	51.89
DFTD	64.03 ± 7.1	73.50 ± 12.0	68.76	64.44 ± 8.7	68.12 ± 16.2	66.28	54.60 ± 2.0	60.19 ± 1.1	57.40
WIKG	57.81 ± 11.5	66.60 ± 8.9	62.20	55.56 ± 6.7	41.98 ± 10.1	48.77	60.56 ± 4.0	61.52 ± 2.1	61.04
ILRA	70.33 ± 7.1	68.82 ± 5.8	69.57	59.01 ± 11.5	68.12 ± 4.5	63.56	63.80 ± 2.5	60.86 ± 1.2	62.33
RRT	68.56 ± 4.1	62.80 ± 6.7	65.68	59.01 ± 6.5	59.01 ± 7.2	59.01	64.78 ± 2.6	63.19 ± 2.3	63.98
CLAM	75.75 ± 4.6	77.17 ± 8.7	76.46	46.67 ± 3.2	46.67 ± 10.2	46.67	57.43 ± 2.6	57.94 ± 2.7	57.68
MeanMIL	66.55 ± 4.8	69.15 ± 7.5	67.85	53.12 ± 3.2	59.01 ± 16.8	56.07	58.56 ± 2.7	60.06 ± 1.6	59.31
QG-MIL	76.62 ± 5.7	71.33 ± 5.2	73.97	59.01 ± 14.7	59.01 ± 11.6	59.01	63.77 ± 6.4	**66.27 ± 5.0**	65.02
QG-MIL Elementwise	79.98 ± 6.1	82.78 ± 4.8	81.38	59.01 ± 8.9	53.12 ± 15.6	56.06	65.20 ± 7.7	65.81 ± 6.2	65.50
QG-MIL no Gate	79.44 ± 4.2	79.69 ± 8.8	79.56	64.44 ± 11.4	**69.51 ± 5.8**	66.97	64.71 ± 5.2	66.13 ± 6.1	65.42
QG-MIL no QKnorm	79.98 ± 7.8	75.93 ± 8.0	77.96	**69.51 ± 14.3**	64.44 ± 10.0	66.97	61.63 ± 9.0	65.82 ± 5.2	63.72
QG-MIL Layernorm	82.67 ± 7.8	75.93 ± 7.3	79.30	59.01 ± 2.4	64.44 ± 12.5	61.72	64.53 ± 10.5	62.53 ± 5.2	63.53
QG-MIL Deep	79.98 ± 7.0	**86.02 ± 4.7**	**83.00**	**69.51 ± 5.1**	**69.51 ± 4.4**	**69.51**	**67.97 ± 6.1**	63.80 ± 9.1	**65.88**
QG-MIL Light	**82.90 ± 8.3**	77.52 ± 4.1	80.21	60.80 ± 12.5	64.44 ± 11.0	62.62	60.95 ± 2.2	63.17 ± 10.2	62.06
Mean Baselines	68.55	70.87	69.71	54.40	52.99	53.70	59.45	59.13	59.29
Mean QG-MIL	**80.22**	**78.46**	**79.34**	**63.04**	**63.50**	**63.27**	**64.11**	**64.79**	**64.45**

QG-MIL improves attention distribution. While the quantitative results underscore the strong predictive performance of the QG-MIL family, the primary motivation behind QG-MIL is to improve the distribution of attention weights in MIL and to enhance robustness against optimization instabilities. In fig. 2, we report the mean top-10% attention mass on the Breast dataset test set. QG-MIL maintains a near-constant mass of approximately 15%, effectively avoiding the attention-sink behavior seen in ABMIL, which concentrates up to 54% of the total attention within the top 10% of instances. To rigorously quantify this mit-

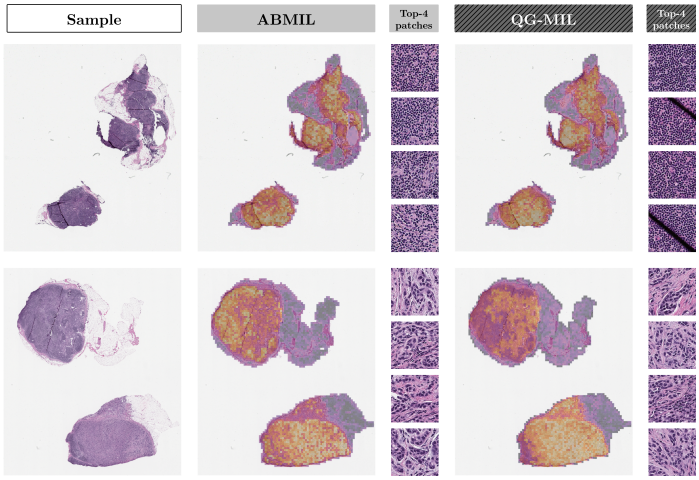

Fig. 3. Qualitative attention overlays on example cases. Left: sample whole slide image. Middle: normalized attention heatmap from ABMIL and Top-4 patches based on attention score. Right: normalized attention heatmap from QG-MIL along the Top-4 patches. QG-MIL highlights a visually smoother attention distribution, avoiding salt-and-pepper-like attention peaks for single patches.

igation, we evaluate the attention distribution using the Gini index and entropy, where lower Gini and higher entropy indicate uniformity. ABMIL exhibits highly concentrated attention (Gini: 0.62 ± 0.13, entropy: 0.90 ± 0.05). In contrast, QG-MIL achieves a significantly more distributed profile (Gini: 0.16 ± 0.07, entropy: 0.99 ± 0.01). This quantitatively supports that QG-MIL's internal architectural gating intrinsically regularizes the pooling distribution without requiring auxiliary entropy losses. This distributed behavior is visually confirmed in fig. 3. Compared to ABMIL, QG-MIL produces smoother, spatially distributed attention maps. Notably, the top-4 attended patches in QG-MIL focus on similar morphological structures rather than isolated high-magnitude peaks, suggesting improved instance coverage and reduced sensitivity to local attention maxima.

4 Conclusion and Limitations

Conclusion and Limitations. We introduce QG-MIL, a domain-agnostic gated transformer aggregation module that structurally mitigates attention sinks in medical imaging MIL without auxiliary losses. Across six pathology and hematology benchmarks, QG-MIL outperformed leading baselines by an average of +6.1 macro F1 points and yielded a smoother, more clinically plausible attention distribution. Despite these advances, our study presents notable limitations. First, the full gated architecture can introduce variance in small patient cohorts, where simpler QG-MIL variants without gating are more effective. Second, while fine-grained gating and SwiGLU modules improve performance, they increase

training memory and time relative to standard aggregators. Importantly, this overhead does not translate to prohibitive inference costs. For realistic bag sizes of 512 and 1024, QG-MIL requires approximately 7 and 18 GFLOPs, respectively, and thus exhibits a lower computational burden than RRT and TransMIL. Finally, exhaustive hyperparameter optimization was omitted to ensure fair baseline comparisons. Future work will explore adaptive gating mechanisms that dynamically scale complexity based on cohort size, alongside expanding evaluations to multi-modal clinical pipelines to further validate QG-MIL's generalizability. **CO2 Emissions from Experiments.** Our experiments were run on our in-house infrastructure using the NVIDIA A100 PCIe 80 GB hardware; the total emissions are estimated at 7.78 kg CO2eq. **Disclosure of Interests** The authors have no competing interests to declare that are relevant to the content of this article. **Acknowledgments** This work was supported by NRRP, Mission 4 Component 2 Investment 1.5, funded by the European Union – NextGenerationEU, project ECS0000038 "eINS Ecosystem of Innovation for Next Generation Sardinia," CUP F53C22000430001. C.M. acknowledges ERC funding under Horizon 2020, Grant Agreements No. 866411 and 101113551, and support from the Hightech Agenda Bayern.

References

1. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
2. Chelebian, E., Avenel, C., Järemo, H., Andersson, P., Wählby, C., Bergh, A.: A clinical prostate biopsy dataset with undetected cancer. Scientific Data **12**(1), 423 (2025). https://doi.org/10.1038/s41597-025-04758-7
3. Chen, R.J., Ding, T., Lu, M.Y., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**, 850–862 (2024). https://doi.org/10.1038/s41591-024-02857-3
4. Dasdelen, M.F., Kukuljan, I., Lienemann, P., Ozlugedik, F., Sadafi, A., Hehr, M., Spiekermann, K., Pohlkamp, C., Marr, C.: Ai-based hematological malignancy prediction from peripheral blood smears in a large diagnostic laboratory cohort. Leukemia **40**(6), 1318–1322 (2026). https://doi.org/10.1038/s41375-026-02934-1
5. Diosdado, J., Gilabert, P., Seguí, S., Borrego, H.: Lunghist700: A dataset of histological images for deep learning in pulmonary pathology. Scientific Data **11**(1), 1088 (2024). https://doi.org/10.1038/s41597-024-03944-3
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015), https://arxiv.org/abs/1512.03385
7. Hehr, M., Sadafi, A., Matek, C., Lienemann, P., Pohlkamp, C., Haferlach, T., Spiekermann, K., Marr, C.: Explainable ai identifies diagnostic cells of genetic aml subtypes. PLOS Digital Health **2** (2023), https://api.semanticscholar.org/CorpusID:257534165
8. Hehr, M., Sadafi, A., Matek, C., Lienemann, P., Pohlkamp, C., Haferlach, T., Spiekermann, K., Marr, C.: Explainable AI identifies diagnostic cells of genetic AML subtypes. PLOS digital health (2023), https://pmc.ncbi.nlm.nih.gov/articles/PMC10016704/

9. Ilse, M., Tomczak, J., Welling, M.: Attention-based Deep Multiple Instance Learning. In: Proceedings of the 35th International Conference on Machine Learning (Jul 2018), https://proceedings.mlr.press/v80/ilse18a.html
10. Koch, V., Wagner, S.J., Kazeminia, S., Sancar, E., Hehr, M., Schnabel, J.A., Peng, T., Marr, C.: DinoBloom: A Foundation Model for Generalizable Cell Embeddings in Hematology . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15012. Springer Nature Switzerland (October 2024)
11. Li, B., Li, Y., Eliceiri, K.W.: Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14313–14323. IEEE, Nashville, TN, USA (Jun 2021). https://doi.org/10.1109/CVPR46437.2021.01409
12. Lu, M.Y., Chen, B., Williamson, D.F.K., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024). https://doi.org/10.1038/s41591-024-02856-4
13. Qiu, Z., Wang, Z., Zheng, B., Huang, Z., Wen, K., Yang, S., Men, R., Yu, L., Huang, F., Huang, S., Liu, D., Zhou, J., Lin, J.: Gated Attention for Large Language Models: Non-linearity, Sparsity, and Attention-Sink-Free (May 2025). https://doi.org/10.48550/arXiv.2505.06708
14. Qu, L., Wang, M., Song, Z., et al.: Bi-directional weakly supervised knowledge distillation for whole slide image classification. Advances in Neural Information Processing Systems **35**, 15368–15381 (2022)
15. Shao, D., Chen, R.J., Song, A.H., Runevic, J., Lu, M.Y., Ding, T., , Mahmood, F.: Do multiple instance learning models transfer? In: International conference on machine learning (2025)
16. Sidhom, J.W., Siddarthan, I.J., Lai, B.S., et al.: Deep learning for diagnosis of acute promyelocytic leukemia via recognition of genomically imprinted morphologic features. npj Precision Oncology **5**, 38 (2021). https://doi.org/10.1038/s41698-021-00179-y
17. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: Multiple Instance Learning Framework with Masked Hard Instance Mining for Whole Slide Image Classification. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4055–4064. IEEE, Paris, France (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.00377
18. Xu, H., Usuyama, N., Bagga, J., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**, 181–188 (2024). https://doi.org/10.1038/s41586-024-07441-w
19. Yixuan, L.: Cost-sensitive multi-kernel elm based on reduced expectation kernel auto-encoder. PLOS ONE **20**(2), 1–20 (02 2025). https://doi.org/10.1371/journal.pone.0314851
20. Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-Challenging Multiple Instance Learning for Whole Slide Image Classification. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LIII. Lecture Notes in Computer Science, vol. 15111, pp. 125–143. Springer (2024). https://doi.org/10.1007/978-3-031-73668-1_8