LEARNING MORPHOLOGICAL EVIDENCE DISTRIBUTIONS FOR WHOLE-SLIDE IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole-slide image (WSI) classification based on multiple-instance learning is weakly supervised because each slide contains thousands of tile instances but typically has only a slide-level label. Standard attention pooling only selects informative instances, yet the pooled continuous feature does not by itself identify the recurring morphology that supports or opposes a diagnosis. To explicitly identify which recurring morphological patterns support or oppose each diagnosis, we introduce *Morphology Evidence Learning via Distributions* (MELD), a multiple-instance model in which an attention module determines which tiles to collect information from, a shared tokenizer maps tile features to distributions over morphology tokens, and class-conditional pointwise mutual information (PMI) assigns signed diagnostic evidence to these tokens for both prediction and interpretation. Across four WSI benchmarks and two frozen feature encoders, MELD achieves the best Macro-F1 on six of eight dataset–encoder settings. Pathologist review shows that tiles assigned to the learned morphology tokens consistently exhibit clinically recognizable morphological patterns. Moreover, projecting the PMI evidence back to individual tiles provides localization that outperforms localization based on the attention maps of representative MIL baselines. Finally, MELD can directly guide rate allocation for WSI compression using tumor-class PMI, assigning higher JPEG quality factors to tiles with positive tumor-class evidence and a low quality factor to all remaining tiles, achieving classification performance comparable to that of uniform JPEG compression with up to a $2.33\times$ reduction in bits per pixel (BPP). Code is available at anonymous.4open.science/MELD.

1 INTRODUCTION

Whole-slide images (WSIs), obtained by digitizing histopathology glass slides into gigapixel-resolution images, have become an important component of modern computational pathology because they preserve fine-grained tissue morphology essential for clinical diagnosis. However, the gigapixel scale that enables such rich morphological characterization also poses substantial challenges for computational analysis. In practice, each WSI is typically divided into non-overlapping tiles for tractable processing, yet only a small fraction of the thousands to tens of thousands of tiles may contain diagnostically relevant morphology. Exhaustively annotating these tiles requires substantial expert effort and is therefore difficult to scale (Campanella et al., 2019). Consequently, most WSI datasets provide only slide-level labels, making weakly supervised learning a natural paradigm for computational pathology.

Multiple-instance learning (MIL) addresses the scarcity of fine-grained annotations by representing each WSI as a bag of tile-level instances and learning a slide-level predictor from slide-level supervision. Early MIL approaches aggregate instances using fixed operators such as mean or max, while modern approaches increasingly adopt learnable attention mechanisms (Wang et al., 2018; Ilse et al., 2018). Attention-based MIL has been particularly effective for WSIs because it enables the model to select and aggregate information from tiles that it considers diagnostically relevant. This selective aggregation is important in gigapixel slides, where a small metastatic focus or localized pathological pattern may otherwise be overwhelmed by thousands of diagnostically irrelevant tiles.

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.

Despite this success, attention alone provides only a partial description of the diagnostic process. We identify three limitations of conventional attention-based MIL methods. **(PI) Attention does not characterize the observed morphology.** An attention weight quantifies how strongly a tile contributes to aggregation, yet it provides no explicit description of the recurring morphological pattern represented by that tile. Consequently, attention alone cannot indicate whether different tiles exhibit the same recurring morphology or recognize which morphological patterns are present in the slide. **(PII) Attention does not assign class-specific diagnostic evidence.** A highly attended tile may contain morphology shared by several classes, favor a competing diagnosis, or become discriminative only through subsequent feature transformations (Shao et al., 2021). Attention magnitude therefore provides no class-specific measure of whether the observed morphology is enriched or depleted under a diagnosis relative to the overall morphology distribution. **(PIII) Prediction and interpretation are defined through different quantities.** Conventional attention-based MIL typically aggregates weighted tile features into a continuous bag embedding and applies a downstream classifier to obtain the final prediction. Interpretation is subsequently derived from attention maps (Ilse et al., 2018; Ye et al., 2026). As a result, the quantities used to explain a prediction are generally distinct from those that determine the class decision, preventing an exact probabilistic account of how the morphology of each tile contributes to the final decision.

These limitations motivate an more interpretable MIL framework that explicitly characterizes *where* diagnostic information is collected, *what* morphology is observed, and *how* the observed morphology contributes to a diagnostic decision. To this end, we propose *Morphology Evidence Learning via Distributions* (MELD), which performs WSI classification in an explicit probability space over learned morphology tokens. MELD consists of three complementary components, an *evidence collection module* that defines where information is collected through attention distributions over tiles, a *morphology characterization module* that describes what is observed using a shared set of trainable morphology codewords, and an *evidence-based classification module* that quantifies how each morphology token supports or opposes each diagnosis. Together, they provide a common probabilistic representation for classification and interpretation.

Specifically, the first two modules jointly transform a WSI into a slide-specific morphology distribution. For each attention branch, MELD defines a probability distribution over tiles and maps every tile representation to a categorical distribution over a shared vocabulary of morphology tokens. Marginalizing the tile-level morphology assignments under the attention distribution then yields a slide-level distribution over morphology tokens. This distribution therefore combines *where* the model collects information with *what* morphology it observes, while organizing recurring tissue patterns through a shared codebook learned solely from slide-level supervision.

The evidence-based classification module then turns this morphology distribution into class-specific diagnostic evidence. MELD assigns each morphology token a pointwise mutual information (PMI) value (Fano & Hawkins, 1961; Jurafsky et al., 2014), defined as the logarithm of the ratio between its probability under a class-conditional morphology prototype and its probability under a reference distribution. Positive and negative PMI values indicate that a morphology token is enriched or depleted under a diagnosis, respectively. By taking the expectation of these token-level evidence values under the slide-specific morphology distribution, MELD obtains the class score directly. Prediction and interpretation are therefore expressed through the same probabilistic quantities, attention determines *where* evidence is collected, morphology assignments determine *what* is observed, and PMI determines *how* the observed morphology supports or opposes each diagnosis. Moreover, we prove that maximizing this evidence score is equivalent to nearest-prototype classification under KL divergence, and that each pairwise class margin admits an exact decomposition into tile-level attention, morphology assignment, and class-conditional log-likelihood-ratio evidence.

We evaluate MELD on CAMELYON16 (Wang et al., 2016), CAMELYON17 (Bandi et al., 2018), BRACS (Brancati et al., 2022), and CPTAC (CPTAC, 2018a;b; Gillette et al., 2020) against 14 representative MIL baselines. MELD achieves strong overall performance across both feature representations, obtaining the best Macro-F1 on six of the eight dataset–encoder settings. In particular, with medical SSL features, MELD improves Macro-F1 on BRACS from 0.673 to 0.709 while achieving 0.951 on CAMELYON16 and 0.945 on CPTAC. Our main contributions are summarized as follows:

- We introduce a morphology tokenizer for MIL that maps tile representations to a shared set of learnable morphology codewords, explicitly characterizing *what* morphology is observed from slide-level supervision.

- We formulate WSI classification through a distributional morphology representation that combines attention-based evidence collection with morphology codeword assignments, yielding slide-specific probability distributions over learned morphology tokens.
- We develop an evidence-based classification rule that assigns signed, class-specific PMI evidence to morphology tokens and directly uses its expectation as the class score. We prove that this rule is exactly equivalent to nearest-prototype classification under KL divergence and derive an exact tile-level decomposition of every pairwise class margin.
- We validate MELD on four WSI benchmarks with two feature encoders, demonstrating strong classification performance, coherent learned morphology codewords, and class-specific diagnostic evidence. Localization and evidence-guided compression further demonstrate the diagnostic and practical utility of the learned evidence.

2 Related Work

MIL-based WSI Classification. WSI classification is commonly formulated as multiple-instance learning (MIL). Methods such as ABMIL (Ilse et al., 2018), CLAM (Lu et al., 2021), DSMIL (Li et al., 2021), and TransMIL (Shao et al., 2021) aggregate informative tiles, while DTFD-MIL (Zhang et al., 2022), MHIM-MIL (Tang et al., 2023), ACMIL (Zhang et al., 2024), and ASMIL (Ye et al., 2026) address pseudo-bag learning, hard-instance mining, attention diversity, and attention instability, respectively. Despite different aggregation mechanisms, these methods typically classify from continuous latent features, without representing morphology or illustrating its diagnostic role.

Prototype- and Concept-based WSI Representation. Prototype-based methods organize recurring tissue morphology into structured representations. PANTHER (Song et al., 2024) discovers morphological patterns largely through task-agnostic clustering or unsupervised representation learning, whereas PTCMIL (Zhao et al., 2026) learns prototypes for each class with downstream supervision. While these approaches provide explicit morphological representations, the learned prototypes generally do not themselves specify how each morphology supports or opposes a particular diagnosis. A complementary line of work introduces semantically defined pathology concepts, often using external knowledge or vision–language models (Zhao et al., 2024; Sun et al., 2025; Qin et al., 2026). In contrast, MELD learns a morphology codebook directly from slide-level supervision and derives signed diagnostic evidence for each morphology token from its class-conditional distribution, without requiring predefined semantic concepts.

Interpretable and Evidence-based MIL. Interpretability has progressed beyond attention maps. Additive MIL (Javed et al., 2022) exchanges the order of attention and classification to enable exact instance-level attribution, and xMIL (Hense et al., 2024) develops post-hoc explanations. PNEA-MIL (Chen et al., 2026) further learns a latent signal supporting a class. MELD differs by defining evidence using PMI at the morphology–class level, which quantifies how recurring morphology supports or against each diagnosis, and the same quantities determine the class score and admit an exact tile-level decomposition. This jointly characterizes *where* evidence is collected, *what* morphology is observed, and *how* it contributes to the diagnosis.

3 Notation

Scalars are denoted by non-bold lowercase letters (e.g., a, β), vectors by bold lowercase letters (e.g., $\boldsymbol{a}$), and matrices by bold uppercase letters (e.g., $\boldsymbol{A}$). Subscripts index the object under consideration. For example, $\boldsymbol{a}_s$ denotes the vector associated with object s, while $\boldsymbol{a}_{sb}$ denotes the vector associated with the pair (s, b). The i-th scalar entry of an indexed vector is selected using square brackets, e.g., $\boldsymbol{a}_s[i]$. We write $[n] = \{1, \ldots, n\}$. For a categorical variable with C possible outcomes, $\boldsymbol{p} \in \Delta^{C-1}$ denotes a C-dimensional probability vector. For two probability vectors $\boldsymbol{p}, \boldsymbol{q} \in \Delta^{C-1}$, the KL divergence is defined as

$$\mathsf{KL}\left(\boldsymbol{p} \,\|\, \boldsymbol{q}\right) = \sum_{c=1}^{C} \boldsymbol{p}[c] \log \frac{\boldsymbol{p}[c]}{\boldsymbol{q}[c]}. \tag{1}$$

We use $\delta(\cdot)$ for the element-wise sigmoid function. We write $\sigma(\cdot)$ for the softmax function:

$$\boldsymbol{\alpha} = \sigma(\boldsymbol{v}), \quad \boldsymbol{\alpha}[i] = \frac{\exp(\boldsymbol{v}[i])}{\sum_{j=1}^{N} \exp(\boldsymbol{v}[j])}, \quad \boldsymbol{v} \in \mathbb{R}^N, \quad \boldsymbol{\alpha} \in \Delta^{N-1}. \tag{2}$$

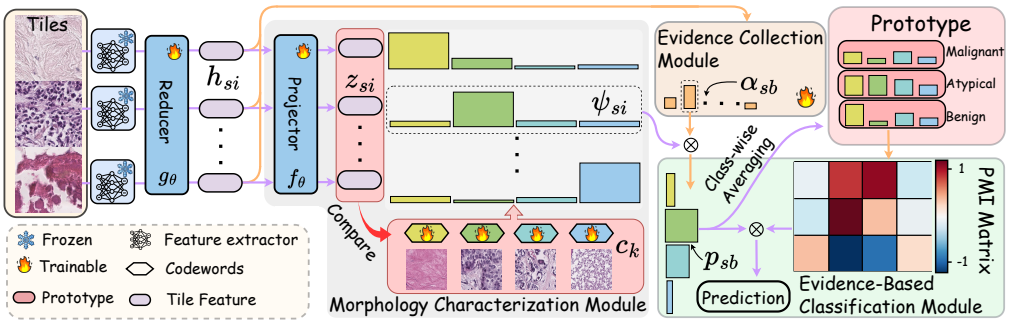

Figure 1: Overview of MELD, which maps frozen tile features to soft morphology-token assignments using a shared tokenizer, while attention defines a normalized distribution over tiles. Marginalizing tile identity under the attention measure yields a slide-level morphology distribution, which is combined with class-specific PMI evidence for prediction. 🔥 and ❄ indicate trainable and frozen components, respectively.

4 METHODOLOGY

We propose *Morphology Evidence Learning via Distributions* (MELD), a multiple-instance learning framework that performs WSI classification in a probability space over learned morphology tokens represented by trainable codewords. Unlike conventional attention-based MIL (Ilse et al., 2018; Shao et al., 2021; Ye et al., 2026), which pools instance features into a continuous slide embedding, MELD constructs a slide-dependent distribution over morphology tokens and classifies the slide by comparing this distribution with empirical class-conditional morphology prototypes.

MELD consists of three modules corresponding to three complementary purposes. The **evidence collection module** determines *where evidence is collected* by assigning an attention distribution over the tiles. The **morphology characterization module** determines *what morphology is observed* by assigning each tile to a shared set of learnable morphology codewords and aggregating these assignments under the attention distribution. Finally, the **evidence-based classification module** determines *how the observed morphology supports or opposes each diagnosis* by assigning every token a pointwise mutual information (PMI) value estimated from the training set. These evidence values are aggregated under the slide-specific morphology distribution to form the class scores, without introducing an additional linear classification head. Formal proofs of the distributional properties and evidence identities used below are provided in Appendix A. Let $X_s = \{\boldsymbol{x}_{si}\}_{i=1}^{N_s}$ denote the features of the s-th WSI, which consists of N_s tiles. These features are extracted by a pretrained image encoder that remains frozen throughout MIL training. Before the MIL model, a trainable feature reducer g_θ, implemented as a linear layer followed by a ReLU activation, maps each tile feature $\boldsymbol{x}_{si}$ to a lower-dimensional representation $\boldsymbol{h}_{si}$.

$$\boldsymbol{h}_{si} = g_\theta(\boldsymbol{x}_{si}) \in \mathbb{R}^D. \tag{3}$$

4.1 EVIDENCE COLLECTION MODULE

A WSI usually contains thousands of tissue tiles, only a subset of which is diagnostically informative. MELD therefore uses B attention branches to construct multiple attention distributions over the same slide. For branch $b \in [B]$, the attention score assigned to tile i is (Ilse et al., 2018)

$$u_{sbi} = \boldsymbol{w}_b^\top \left[\tanh\left(\boldsymbol{V}_b \boldsymbol{h}_{si}\right) \odot \delta\left(\boldsymbol{U}_b \boldsymbol{h}_{si}\right)\right], \tag{4}$$

where $\boldsymbol{w}_b$, $\boldsymbol{V}_b$, and $\boldsymbol{U}_b$ are trainable parameters. Normalizing across all tiles gives

$$\boldsymbol{\alpha}_{sb} = \sigma(\boldsymbol{u}_{sb}), \qquad \boldsymbol{\alpha}_{sb}[i] = \frac{\exp(u_{sbi})}{\sum_{j=1}^{N_s} \exp(u_{sbj})}, \qquad \boldsymbol{\alpha}_{sb} \in \Delta^{N_s-1}. \tag{5}$$

Hence, each branch defines a slide-dependent probability distribution over tile indices. The value $\boldsymbol{\alpha}_{sb}[i]$ specifies how much attention mass branch b assigns to tile i, thereby determining *where* the subsequent morphology evidence is collected and controlling the contribution of each tile to the downstream morphology distribution. Following Zhang et al. (2024), to discourage different

branches from repeatedly selecting the same regions, we use the attention-diversity regularizer

$$\mathcal{L}_{\text{div}}^{(s)} = \frac{2}{B(B-1)} \sum_{b<b'} \frac{\boldsymbol{\alpha}_{sb}^{\top} \boldsymbol{\alpha}_{sb'}}{\|\boldsymbol{\alpha}_{sb}\|_2 \|\boldsymbol{\alpha}_{sb'}\|_2}. \tag{6}$$

Minimizing Equation 6 encourages diversity among the attention distributions across branches.

4.2 Morphology Characterization Module

The evidence collection module determines which tiles contribute most strongly to the slide representation, but does not explicitly characterize the morphology represented by those tiles. To characterize the observed morphology, the tokenizer first maps the reduced feature into a shared morphology embedding space using a trainable projector f_θ

$$\boldsymbol{z}_{si} = f_\theta(\boldsymbol{h}_{si}) \in \mathbb{R}^{D_q}, \tag{7}$$

where $\boldsymbol{z}_{si}$ denotes the morphology embedding of tile i from slide s. Within this embedding space, MELD learns a shared codebook $\mathcal{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K\}$, where each $\boldsymbol{c}_k \in \mathbb{R}^{D_q}$ is a trainable morphology codeword. We use morphology token k to denote the categorical outcome represented by codeword $\boldsymbol{c}_k$. To assign each tile to the learned morphology codewords, we compute the cosine similarity between the normalized representation and the normalized codewords.

$$\boldsymbol{\psi}_{si}[k] = \frac{\exp\left(\overline{\boldsymbol{z}}_{si}^{\top} \overline{\boldsymbol{c}}_k / \gamma_q\right)}{\sum_{\ell=1}^{K} \exp\left(\overline{\boldsymbol{z}}_{si}^{\top} \overline{\boldsymbol{c}}_\ell / \gamma_q\right)}, \quad \overline{\boldsymbol{z}}_{si} = \frac{\boldsymbol{z}_{si}}{\|\boldsymbol{z}_{si}\|_2}, \quad \overline{\boldsymbol{c}}_k = \frac{\boldsymbol{c}_k}{\|\boldsymbol{c}_k\|_2}, \qquad \boldsymbol{\psi}_{si} \in \Delta^{K-1}, \tag{8}$$

where $\gamma_q > 0$ is a temperature parameter controlling the concentration of the distribution. Equation 8 defines a differentiable soft assignment over morphology tokens. As γ_q decreases, the distribution becomes increasingly concentrated on the codeword with the highest cosine similarity, whereas a non-zero temperature preserves differentiability and allows both the projector and the codebook to be optimized jointly with the MIL model. While $\boldsymbol{\psi}_{si}$ characterizes the morphology of an individual tile, MELD represents each slide by aggregating these tile-level morphology distributions according to the attention assigned by each branch:

$$\boldsymbol{p}_{sb} = \sum_{i=1}^{N_s} \boldsymbol{\alpha}_{sb}[i] \boldsymbol{\psi}_{si}, \qquad \boldsymbol{p}_{sb} \in \Delta^{K-1}. \tag{9}$$

Accordingly, $\boldsymbol{p}_{sb}[k]$ denotes the attention-weighted probability mass assigned to morphology token k by branch b for slide s. We visualize the token assignment in Figure 3 and Appendix F.

4.3 Evidence-Based Classification Module

The morphology characterization module describes which morphology tokens are observed, but does not yet specify how they support or oppose each diagnosis. MELD estimates this class-specific diagnostic evidence empirically from the training set. For each class y, let $\mathcal{D}_y = \{s : y_s = y\}$ denote the corresponding training slides. The branch-specific class-conditional morphology prototype is estimated as

$$\boldsymbol{q}_{yb} = \frac{1}{|\mathcal{D}_y|} \sum_{s \in \mathcal{D}_y} \boldsymbol{p}_{sb}, \qquad \boldsymbol{q}_{yb} \in \Delta^{K-1}. \tag{10}$$

To define a class-balanced reference distribution for interpreting morphology enrichment and depletion, we use a class-balanced prior $P_b(y) = 1/C$. This gives the branch-specific joint distribution and morphology marginal

$$P_b(y, k) = P_b(y) P_b(k \mid y) = \frac{1}{C} \boldsymbol{q}_{yb}[k], \quad P_b(k) = \boldsymbol{m}_b[k] = \frac{1}{C} \sum_{c=1}^{C} \boldsymbol{q}_{cb}[k]. \tag{11}$$

The diagnostic evidence associated with morphology token k for class y in branch b is given by its pointwise mutual information (PMI),

$$\boldsymbol{e}_{yb}[k] = \log \frac{P_b(y, k)}{P_b(y) P_b(k)} = \log \frac{\boldsymbol{q}_{yb}[k]}{\boldsymbol{m}_b[k]}. \tag{12}$$

Thus, positive PMI indicates that morphology token k is enriched under class y relative to the class-agnostic marginal, whereas negative PMI indicates that it is depleted. Values near zero indicate little class-specific deviation from the reference distribution. MELD directly aggregates these evidence values under the slide-specific morphology distributions to form the class score:

$$s_y(X_s) = \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} \boldsymbol{p}_{sb}[k]\boldsymbol{e}_{yb}[k]. \tag{13}$$

Hence, the expected PMI directly serves as the prediction score.

Remark 1 (Equivalent Decision Rule)**.** *The PMI evidence score admits an equivalent interpretation in KL geometry:*

$$\arg\max_{y\in[C]} s_y(X_s) = \arg\min_{y\in[C]} \frac{1}{B}\sum_{b=1}^{B} \mathsf{KL}(\boldsymbol{p}_{sb} \,\|\, \boldsymbol{q}_{yb})\,. \tag{14}$$

Hence, maximizing signed PMI evidence is exactly equivalent to nearest-prototype classification under KL divergence. The derivation is provided in Appendix A.2.

MELD applies softmax to the evidence scores and is trained with standard cross-entropy. By the equivalent decision rule above, this is also equivalent to cross-entropy classification based on the negative average KL divergence between the slide-specific morphology distributions and the class prototypes. Importantly, the same quantities used for prediction also yield an exact tile-level decomposition. Define the contribution of tile i as

$$\tau_{si}^{y} = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{\alpha}_{sb}[i]\sum_{k=1}^{K} \boldsymbol{\psi}_{si}[k]\boldsymbol{e}_{yb}[k], \qquad s_y(X_s) = \sum_{i=1}^{N_s} \tau_{si}^{y}. \tag{15}$$

Each tile contribution therefore combines the attention weight determining *where* evidence is collected, the morphology assignment describing *what* is observed, and the class-specific PMI specifying *how* that morphology supports or opposes the diagnosis. Moreover, the pairwise margin between any two classes $c, r \in [C]$ also decomposes exactly over tiles:

$$s_c(X_s) - s_r(X_s) = \sum_{i=1}^{N_s} (\tau_{si}^{c} - \tau_{si}^{r})\,, \quad \tau_{si}^{c} - \tau_{si}^{r} = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{\alpha}_{sb}[i]\sum_{k=1}^{K} \boldsymbol{\psi}_{si}[k]\log\frac{\boldsymbol{q}_{cb}[k]}{\boldsymbol{q}_{rb}[k]}. \tag{16}$$

Thus, the preference for class c over class r is exactly the sum of tile-level contributions, where each morphology token contributes through its class-conditional log-likelihood ratio. Complete derivations of the score, class-margin, and predictive log-odds decompositions are provided in Appendix A.3. We further evaluate their faithfulness through deletion interventions in Appendix D.

5 Experiments

We organize our experiments around three questions. **(I)** How well does MELD perform on whole-slide classification? We compare MELD with representative MIL methods across four datasets and two encoders. **(II)** Do MELD's components work as intended? We evaluate the three modules according to their distinct roles: whether attention covers diagnostically relevant regions, whether the tokenizer organizes coherent morphology, and whether the Evidence-Based Classification Module assigns diagnostically meaningful signed evidence. **(III)** Can the learned evidence be used beyond interpretation and classification? We investigate PMI-guided rate allocation for WSI compression, where diagnostically important regions are preserved at higher quality while less important regions are more aggressively compressed, substantially reducing bits per pixel (BPP) while maintaining downstream classification accuracy.

Datasets and experimental setup. We evaluate MELD on four WSI benchmarks: CAMELYON16 (Ehteshami Bejnordi et al., 2017), CAMELYON17 (Bandi et al., 2018), BRACS (Brancati et al., 2022), and CPTAC (CPTAC, 2018a;b). In particular, CAMELYON17 is evaluated as a four-class classification task using a center-wise split, allowing us to evaluate cross-center generalization under acquisition conditions that differ from those represented during training. We use ImageNet-pretrained (Deng et al., 2009) ResNet-18 (He et al., 2016) and self-supervised ViT (Caron et al.,

Table 1: The Macro-$F1$ and Macro-AUROC of different MIL approaches across four WSI datasets. **Bold** and underlined values denote the best and second-best results, respectively.

Backbone	Dataset	CAMELYON16		CAMELYON17		BRACS		CPTAC	
	Method	$F1\uparrow$	AUROC $\uparrow$	$F1\uparrow$	AUROC $\uparrow$	$F1\uparrow$	AUROC $\uparrow$	$F1\uparrow$	AUROC $\uparrow$
ResNet-18 ImageNet-pretrained	ABMIL ICML 2018	$0.777_{\pm 0.004}$	$0.795_{\pm 0.018}$	$0.479_{\pm 0.043}$	$0.766_{\pm 0.040}$	$0.501_{\pm 0.018}$	$0.694_{\pm 0.009}$	$0.886_{\pm 0.027}$	$0.953_{\pm 0.013}$
	CLAM-SB Nature 2021	$0.766_{\pm 0.029}$	$0.806_{\pm 0.038}$	$0.500_{\pm 0.030}$	$0.785_{\pm 0.060}$	$0.504_{\pm 0.038}$	$0.749_{\pm 0.045}$	$0.887_{\pm 0.012}$	$0.954_{\pm 0.003}$
	DSMIL CVPR 2021	$0.725_{\pm 0.022}$	$0.779_{\pm 0.057}$	$0.403_{\pm 0.035}$	$0.699_{\pm 0.046}$	$0.511_{\pm 0.037}$	$0.759_{\pm 0.015}$	$0.860_{\pm 0.010}$	$0.936_{\pm 0.006}$
	TransMIL NeurIPS 2021	$0.713_{\pm 0.072}$	$0.768_{\pm 0.059}$	$0.521_{\pm 0.061}$	$0.771_{\pm 0.037}$	$0.509_{\pm 0.033}$	$0.753_{\pm 0.018}$	$0.886_{\pm 0.013}$	$0.937_{\pm 0.006}$
	DTFD-MIL CVPR 2022	$\underline{0.783}_{\pm 0.003}$	$0.791_{\pm 0.027}$	$0.512_{\pm 0.037}$	$0.770_{\pm 0.038}$	$0.465_{\pm 0.068}$	$0.688_{\pm 0.070}$	$0.878_{\pm 0.016}$	$0.950_{\pm 0.007}$
	Additive MIL NeurIPS 2022	$0.729_{\pm 0.103}$	$0.782_{\pm 0.099}$	$0.486_{\pm 0.033}$	$0.779_{\pm 0.026}$	$0.487_{\pm 0.052}$	$0.718_{\pm 0.018}$	$0.888_{\pm 0.011}$	$0.959_{\pm 0.005}$
	IBMIL CVPR 2023	$0.776_{\pm 0.025}$	$0.810_{\pm 0.033}$	$0.483_{\pm 0.048}$	$0.798_{\pm 0.033}$	$0.514_{\pm 0.011}$	$0.755_{\pm 0.043}$	$0.903_{\pm 0.006}$	$0.957_{\pm 0.005}$
	MHIM-MIL ICCV 2023	$0.778_{\pm 0.020}$	$0.796_{\pm 0.034}$	$0.462_{\pm 0.046}$	$0.765_{\pm 0.040}$	$\mathbf{0.530}_{\pm 0.035}$	$\mathbf{0.788}_{\pm 0.020}$	$0.880_{\pm 0.007}$	$0.960_{\pm 0.004}$
	ACMIL ECCV 2024	$0.766_{\pm 0.023}$	$0.780_{\pm 0.026}$	$0.495_{\pm 0.041}$	$0.790_{\pm 0.038}$	$0.478_{\pm 0.044}$	$0.759_{\pm 0.017}$	$0.894_{\pm 0.007}$	$0.959_{\pm 0.004}$
	CAMIL ICLR 2024	$0.705_{\pm 0.028}$	$0.742_{\pm 0.053}$	$0.518_{\pm 0.054}$	$0.794_{\pm 0.044}$	$0.495_{\pm 0.052}$	$0.716_{\pm 0.017}$	$0.887_{\pm 0.010}$	$0.953_{\pm 0.004}$
	PANTHER CVPR 2024	$0.710_{\pm 0.049}$	$0.770_{\pm 0.046}$	$0.455_{\pm 0.049}$	$0.751_{\pm 0.028}$	$0.469_{\pm 0.036}$	$0.642_{\pm 0.022}$	$0.877_{\pm 0.006}$	$0.913_{\pm 0.009}$
	PTCMIL MIA 2026	$0.782_{\pm 0.021}$	$0.810_{\pm 0.057}$	$0.516_{\pm 0.043}$	$0.799_{\pm 0.038}$	$0.524_{\pm 0.077}$	$0.752_{\pm 0.010}$	$0.895_{\pm 0.016}$	$0.956_{\pm 0.003}$
	ASMIL ICLR 2026	$0.743_{\pm 0.006}$	$0.735_{\pm 0.024}$	$\underline{0.522}_{\pm 0.014}$	$\underline{0.800}_{\pm 0.022}$	$0.519_{\pm 0.032}$	$0.734_{\pm 0.019}$	$0.887_{\pm 0.012}$	$0.963_{\pm 0.008}$
	PNEA-MIL ACM BCB 2026	$0.766_{\pm 0.017}$	$\underline{0.810}_{\pm 0.013}$	$0.423_{\pm 0.006}$	$0.727_{\pm 0.009}$	$0.452_{\pm 0.012}$	$0.747_{\pm 0.008}$	$\underline{0.907}_{\pm 0.003}$	$\mathbf{0.973}_{\pm 0.002}$
	MELD	$\mathbf{0.790}_{\pm 0.008}$	$\mathbf{0.820}_{\pm 0.023}$	$\mathbf{0.537}_{\pm 0.013}$	$\mathbf{0.828}_{\pm 0.012}$	$\underline{0.529}_{\pm 0.043}$	$\underline{0.764}_{\pm 0.023}$	$\mathbf{0.914}_{\pm 0.011}$	$\underline{0.965}_{\pm 0.009}$
ViT-S SSL-pretrained	ABMIL ICML 2018	$0.920_{\pm 0.030}$	$0.943_{\pm 0.022}$	$0.614_{\pm 0.090}$	$0.829_{\pm 0.035}$	$0.648_{\pm 0.066}$	$0.858_{\pm 0.003}$	$0.919_{\pm 0.005}$	$0.976_{\pm 0.003}$
	CLAM-SB Nature 2021	$0.925_{\pm 0.049}$	$0.952_{\pm 0.046}$	$0.543_{\pm 0.047}$	$0.856_{\pm 0.017}$	$0.604_{\pm 0.046}$	$0.851_{\pm 0.008}$	$0.926_{\pm 0.006}$	$0.976_{\pm 0.005}$
	DSMIL CVPR 2021	$0.924_{\pm 0.036}$	$0.964_{\pm 0.016}$	$0.556_{\pm 0.099}$	$0.820_{\pm 0.046}$	$0.571_{\pm 0.053}$	$0.826_{\pm 0.022}$	$0.908_{\pm 0.016}$	$0.974_{\pm 0.003}$
	TransMIL NeurIPS 2021	$0.921_{\pm 0.011}$	$0.949_{\pm 0.023}$	$0.623_{\pm 0.064}$	$0.869_{\pm 0.049}$	$0.613_{\pm 0.040}$	$0.838_{\pm 0.013}$	$0.919_{\pm 0.009}$	$0.970_{\pm 0.004}$
	DTFD-MIL CVPR 2022	$\underline{0.942}_{\pm 0.007}$	$0.973_{\pm 0.015}$	$0.638_{\pm 0.062}$	$\underline{0.902}_{\pm 0.032}$	$0.594_{\pm 0.038}$	$0.866_{\pm 0.029}$	$0.923_{\pm 0.014}$	$0.977_{\pm 0.003}$
	Additive MIL NeurIPS 2022	$0.936_{\pm 0.020}$	$\underline{0.980}_{\pm 0.005}$	$0.582_{\pm 0.053}$	$0.866_{\pm 0.013}$	$0.562_{\pm 0.058}$	$0.796_{\pm 0.006}$	$0.917_{\pm 0.014}$	$0.978_{\pm 0.003}$
	IBMIL CVPR 2023	$0.910_{\pm 0.019}$	$0.944_{\pm 0.028}$	$0.593_{\pm 0.063}$	$0.860_{\pm 0.039}$	$0.664_{\pm 0.044}$	$0.858_{\pm 0.011}$	$0.921_{\pm 0.008}$	$0.971_{\pm 0.005}$
	MHIM-MIL ICCV 2023	$0.906_{\pm 0.014}$	$0.950_{\pm 0.018}$	$0.593_{\pm 0.073}$	$0.855_{\pm 0.046}$	$0.641_{\pm 0.063}$	$\underline{0.871}_{\pm 0.004}$	$0.895_{\pm 0.019}$	$0.973_{\pm 0.005}$
	ACMIL ECCV 2024	$0.903_{\pm 0.067}$	$0.969_{\pm 0.007}$	$0.613_{\pm 0.031}$	$\mathbf{0.906}_{\pm 0.029}$	$0.602_{\pm 0.059}$	$0.869_{\pm 0.011}$	$0.924_{\pm 0.008}$	$0.978_{\pm 0.002}$
	CAMIL ICLR 2024	$0.936_{\pm 0.010}$	$0.960_{\pm 0.018}$	$\mathbf{0.662}_{\pm 0.068}$	$0.871_{\pm 0.023}$	$\underline{0.673}_{\pm 0.025}$	$0.836_{\pm 0.028}$	$0.913_{\pm 0.005}$	$0.963_{\pm 0.003}$
	PANTHER CVPR 2024	$0.906_{\pm 0.061}$	$0.956_{\pm 0.074}$	$0.576_{\pm 0.046}$	$0.816_{\pm 0.036}$	$0.628_{\pm 0.019}$	$0.794_{\pm 0.014}$	$0.901_{\pm 0.008}$	$0.942_{\pm 0.009}$
	PTCMIL MIA 2026	$0.938_{\pm 0.010}$	$0.973_{\pm 0.012}$	$0.650_{\pm 0.046}$	$0.898_{\pm 0.041}$	$0.629_{\pm 0.046}$	$0.856_{\pm 0.016}$	$0.923_{\pm 0.006}$	$0.975_{\pm 0.004}$
	ASMIL ICLR 2026	$0.940_{\pm 0.003}$	$0.976_{\pm 0.004}$	$0.636_{\pm 0.061}$	$0.848_{\pm 0.050}$	$0.667_{\pm 0.051}$	$0.824_{\pm 0.017}$	$0.920_{\pm 0.015}$	$0.970_{\pm 0.004}$
	PNEA-MIL ACM BCB 2026	$0.904_{\pm 0.025}$	$0.965_{\pm 0.009}$	$0.595_{\pm 0.031}$	$0.848_{\pm 0.023}$	$0.480_{\pm 0.034}$	$0.808_{\pm 0.008}$	$\underline{0.927}_{\pm 0.003}$	$\underline{0.982}_{\pm 0.001}$
	MELD	$\mathbf{0.951}_{\pm 0.011}$	$\mathbf{0.983}_{\pm 0.003}$	$\underline{0.652}_{\pm 0.045}$	$0.896_{\pm 0.024}$	$\mathbf{0.709}_{\pm 0.020}$	$\mathbf{0.878}_{\pm 0.011}$	$\mathbf{0.945}_{\pm 0.008}$	$\mathbf{0.983}_{\pm 0.001}$

2021; Kang et al., 2023) as frozen feature encoders. For all classification experiments, we use the same five random seeds, $\{1, 2, 3, 4, 5\}$, and all baselines are re-run under the same computational environment for controlled comparison. We report the mean and standard deviation of Macro-$F1$ and Macro-AUROC across the five runs[1]. All methods use the same PyTorch (Paszke et al., 2019) implementation of these metrics, with single-label predictions obtained by $\arg\max$ over class probabilities. Complete training, model-selection, prototype-refresh, dataset splits, and statistical evaluation protocols are provided in Appendix B.

5.1 Whole-Slide Classification

We first evaluate the slide-level classification performance of MELD. Table 1 compares MELD with representative MIL methods on four WSI datasets using two frozen feature encoders. MELD achieves consistently strong performance across datasets and feature representations. With ViT-S features, MELD achieves the best Macro-$F1$ and Macro-AUROC on CAMELYON16, BRACS, and CPTAC, obtaining $0.951/0.983$, $0.709/0.878$, and $0.945/0.983$, respectively. The improvement is particularly pronounced on BRACS, where MELD increases Macro-$F1$ from 0.673 for the strongest competing method to 0.709. On the cross-center CAMELYON17 benchmark, MELD obtains $0.652/0.896$, remaining competitive although not outperforming the strongest baselines. With ResNet-18 features, MELD achieves the best Macro-$F1$ and Macro-AUROC on both CAMELYON16 and CAMELYON17, reaching $0.790/0.820$ and $0.537/0.828$, respectively. It further ranks second on both metrics on BRACS and achieves the best Macro-$F1$ on CPTAC. Overall, MELD shows strong performance across datasets and feature representations.

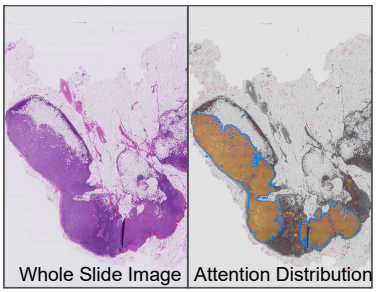

Figure 2: **Left:** the original whole-slide image. **Right:** the attention distribution averaged across branches, overlaid on the WSI, with blue contours indicating the tumor ROI. The attention covers most of the tumor region.

5.2 Effectiveness of MELD Components

We next examine whether the three modules of MELD produce the intended representations. We evaluate the Evidence Collection Module based on whether its attention covers diagnostically relevant regions, the Morphology Characterization Module by the semantic coherence of its learned

[1] See Appendix C for details on metric computation and interpretation.

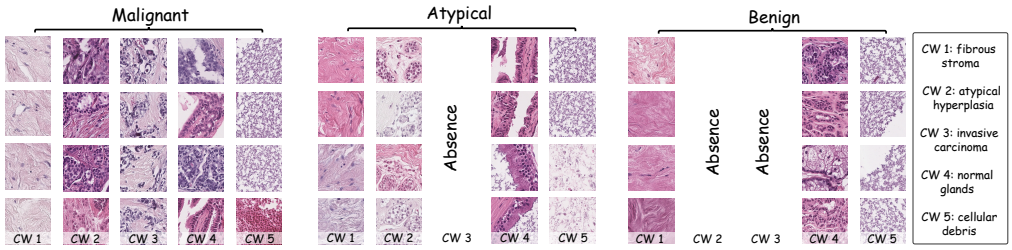

Figure 3: Visualization of the codewords learned by MELD on BRACS. A pathologist identifies consistent morphology across the learned codewords: Codeword 1 is associated with fibrous stroma, Codeword 2 with atypical hyperplasia, Codeword 3 with invasive carcinoma, Codeword 4 with normal glands, and Codeword 5 with cellular debris. For each diagnostic class and codeword with assigned tiles, we randomly sample eight tiles whose highest-probability assignment is to that codeword for pathologist review, and randomly select four of the eight reviewed tiles for visualization.

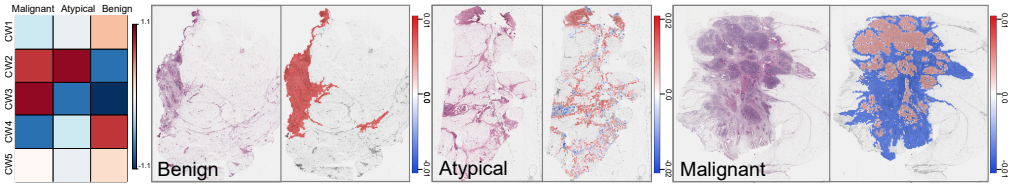

Figure 4: Mechanistic visualization of the Evidence-Based Classification Module on BRACS. The PMI matrix captures class-specific morphology associations: positive values indicate enrichment relative to the class-balanced marginal, while negative values indicate depletion. To isolate class-specific morphology evidence from attention-based evidence collection, we visualize the attention-decoupled score in Equation (18). Red regions indicate tiles that provide positive evidence for the corresponding diagnosis, whereas blue regions indicate tiles that provide negative evidence. For benign slides, the tissue predominantly supports the benign decision. For atypical and malignant slides, MELD successfully highlights atypical and cancerous regions as positive evidence, while normal regions provide negative evidence against these diagnoses.

codewords, and the Evidence-Based Classification Module by whether its signed PMI evidence agrees with pathological interpretation. Ablations of design choices are provided in Appendix E.

Evidence Collection Module. The Evidence Collection Module determines where MELD collects information for subsequent morphology characterization and diagnostic reasoning. We evaluate whether its learned attention covers diagnostically relevant regions on CAMELYON16, which provides pixel-level tumor annotations. Specifically, we average the attention distributions across all branches and rescale the resulting attention map by the number of tissue tiles to account for differences in bag size. We then obtain a binary collection mask by selecting tiles whose rescaled attention exceeds 1, and compute recall with respect to the annotated tumor ROI, measuring the fraction of tumor tissue captured by the module. MELD achieves a recall of 0.94 and a Dice score of 0.34 with ImageNet-pretrained features, and a recall of 0.96 and a Dice score of 0.37 with medical self-supervised (SSL) features, indicating that the Evidence Collection Module covers nearly all annotated tumor regions while maintaining meaningful spatial overlap under both feature representations. Figure 2 provides a randomly selected example from the test set, where the attention closely covers the tumor ROI. We verify the effectiveness of the Evidence Collection Module in Appendix E.1.

Morphology Characterization Module. We next examine whether the learned codewords capture coherent and recognizable tissue morphology. For pathology analysis, we use a compact $K = 5$ morphology vocabulary on BRACS, allowing all learned codewords to be inspected individually. For each codeword, we randomly sample eight assigned tiles for pathologist review. After the pathologist confirms their morphological consistency and identifies the associated morphology, four of the eight reviewed tiles are randomly selected for visualization. As shown in Figure 3, each codeword corresponds to a consistent and clinically recognizable pattern, and the assignments are also class-dependent. We notice that Codeword 3 (invasive carcinoma) appears only in malignant cases, while Codeword 2 (hyperplasia) appears in malignant and atypical cases but is absent from benign ones. These observations suggest that the Morphology Characterization Module organizes WSI tiles into reusable morphology tokens that correspond to clinically meaningful tissue patterns, despite being trained using only slide-level supervision. Additional qualitative visualizations using separately trained MELD models with a larger morphology vocabulary are provided in Appendix F.

Evidence-Based Classification Module. We examine whether MELD learns diagnostically meaningful evidence and whether these learned morphology–class associations can be spatially mapped back to tissue regions. For visualization, we first average the PMI across branches,

$$\bar{e}_y[k] = \frac{1}{B}\sum_{b=1}^{B} e_{yb}[k]. \tag{17}$$

Figure 4 visualizes the resulting branch-averaged PMI matrix, where positive and negative values indicate that a morphology token supports or opposes a diagnosis, respectively. To examine the Evidence-Based Classification Module independently of attention-based evidence collection, we omit the attention weights and define the *morphology evidence* of tile i for class y as

$$r_{si}^{y} = \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K} \psi_{si}[k] e_{yb}[k]. \tag{18}$$

Thus, r_{si}^{y} measures the class-specific association of the local morphology with class y, independently of attention-based evidence collection. Mapping r_{si}^{y} spatially yields the evidence maps in Figure 4, where positive and negative values indicate enrichment and depletion of local morphology, respectively, with respect to class y relative to the reference distribution. On CAMELYON16, this localization outperforms attention-based MIL baselines; quantitative evaluation and additional qualitative results are provided in Appendix G.

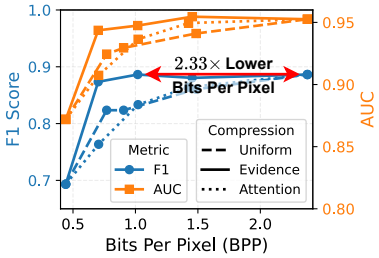

Figure 5: Rate–accuracy trade-off of WSI compression. Evidence-guided compression achieves the best F1 score and AUROC at comparable average BPP, outperforming both uniform and attention-guided compression. Attention-guided compression is less effective and can underperform uniform compression at low rates.

5.3 Evidence-Guided WSI Compression

We finally examine whether the diagnostic evidence learned by MELD can guide rate allocation for WSI compression on CAMELYON16. We first train MELD using frozen UNI (Chen et al., 2024) features and compute the morphology evidence for the positive tumor class at each tile. Tiles with positive tumor-class evidence are preserved at a higher JPEG quality factor, whereas the remaining tiles are compressed aggressively using a fixed low quality factor of 10. We vary the quality factor assigned to positive-evidence tiles to obtain different operating points along the rate–accuracy curve. We further consider an attention-guided strategy that preserves the same number of tiles as the evidence-guided strategy, selecting those with the highest attention scores and assigning them the same higher quality factor.

To evaluate whether the retained information remains useful to an independent MIL classifier, we evaluate the compressed WSIs using DSMIL (Li et al., 2021) with UNI features. We compare both adaptive strategies with uniform JPEG compression, in which all tiles are encoded using the same quality factor. Figure 5 reports Macro-$F1$ and Macro-AUROC as functions of the average bits per pixel (BPP). Evidence-guided compression consistently achieves the best rate–accuracy trade-off, requiring up to 2.33× lower BPP than uniform compression to achieve the same Macro-$F1$. In contrast, attention-guided compression is less effective and can even underperform uniform compression at low rates. These results suggest that MELD's class-specific diagnostic evidence provides a more informative prior for rate allocation than attention alone.

6 Conclusion

We introduced MELD, a novel interpretable MIL framework for weakly supervised WSI classification that models where diagnostic evidence is collected, what morphology is observed, and how the observed morphology supports or opposes each diagnosis. MELD represents each slide as a morphology-token distribution and uses signed, class-specific PMI evidence for prediction and interpretation. We further showed that the resulting class score is exactly equivalent to KL-based prototype classification and that each pairwise class margin admits an exact tile-level decomposition. Experiments across four WSI benchmarks demonstrate strong classification performance, clinically recognizable morphology tokens, class-specific localization, and practical utility for evidence-guided compression. Due to space limitations, we defer discussion of the limitations to Appendix H.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation. The authors reviewed the resulting text, equations, and claims, verified the final technical content, and took responsibility for the methodology, experiments, results, citations, and conclusions reported in the manuscript.

ETHICS STATEMENT

All WSI datasets used in this work are publicly available and were obtained from open-access websites. The usage of these datasets strictly follows the terms and conditions set by the dataset providers and adheres to established academic and research community standards. Only de-identified public data were used; no direct personal identifiers were accessed.

ACKNOWLEDGMENT

Data used in this publication were obtained from the CAMELYON16 challenge dataset. Data used in this publication were obtained from the CAMELYON17 challenge dataset. Data used in this publication were obtained from the BReAst Carcinoma Subtyping (BRACS) dataset. Data used in this publication were generated by the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC).

REPRODUCIBILITY STATEMENT

We have taken steps to ensure our results are reproducible. We provide implementation and training details, including dataset task definitions, prototype initialization and refresh, optimization, checkpoint selection, and test evaluation, in Appendix B. Definitions of the evaluation metrics are provided in Appendix C. An anonymized repository at `https://anonymous.4open.science/MELD` contains the source code. All datasets used in our experiments are publicly available.

REFERENCES

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosue Scognamiglio, Guillaume Jaume, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022:baac093, 2022.

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.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Junxiang Chen, Justin L Couetil, Nigel Gordon Maher, Richard A Scolyer, Ahmed Alomari, Kun Huang, and Jie Zhang. Pnea-mil: Interpretable multiple-instance learning for whole-slide images through positive-negative evidence analysis. In *Proceedings of the 17th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–6, 2026.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

CPTAC. The clinical proteomic tumor analysis consortium lung squamous cell carcinoma collection (cptac-lscc), 2018a. URL https://doi.org/10.7937/K9/TCIA.2018.6EMUB5L2. Version 15.

CPTAC. The clinical proteomic tumor analysis consortium lung adenocarcinoma collection (cptac-luad), 2018b. URL https://doi.org/10.7937/K9/TCIA.2018.PAT12TBS.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

Robert M Fano and David Hawkins. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794, 1961.

Olga Fourkioti, Matt De Vries, and Chris Bakal. Camil: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images. In *International Conference on Learning Representations*, 2024.

Michael A Gillette, Shankha Satpathy, Song Cao, Saravana M Dhanasekaran, Suhas V Vasaikar, Karsten Krug, Francesca Petralia, Yize Li, Wen-Wei Liang, Boris Reva, et al. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell*, 182(1):200–225, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Julius Hense, Mina Jamshidi Idaji, Oliver Eberle, Thomas Schnake, Jonas Dippel, Laure Ciernik, Oliver Buchstab, Andreas Mock, Frederick Klauschen, and Klaus-Robert Müller. xmil: Insightful explanations for multiple instance learning in histopathology. *Advances in Neural Information Processing Systems*, 37:8300–8328, 2024.

Maximilian Ilse, Jakub M Tomczak, and Max Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.

Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: Intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022.

Dan Jurafsky, James H Martin, et al. *Speech and language processing*, volume 3. Pearson London, 2014.

Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3344–3354. IEEE, 2023.

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 (CVPR)*, pp. 14318–14328, June 2021.

Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19830–19839, 2023.

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*, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Weiyi Qin, Yingci Liu-Swetz, Shiwei Tan, and Hao Wang. Clear-hpv: Interpretable concept discovery for human-papillomavirus-associated morphology in whole-slide histology. *Patterns*, 2026.

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.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Susu Sun, Dominique van Midden, Geert Litjens, and Christian F Baumgartner. Prototype-based multiple instance learning for gigapixel whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 507–517. Springer, 2025.

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 *IEEE/CVF International Conference on Computer Vision*, 2023.

Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016.

Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern recognition*, 74:15–24, 2018.

Linfeng Ye, Shayan Mohajer Hamidi, Zhixiang Chi, Guang Li, Mert Pilanci, Takahiro Ogawa, Miki Haseyama, and Konstantinos N. Plataniotis. ASMIL: Attention-stabilized multiple instance learning for whole-slide imaging. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CYmjrbQRyM.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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.

Beidi Zhao, Sangmook Kim, Hao Chen, Chen Zhou, Zu-Hua Gao, Gang Wang, and Xiaoxiao Li. Ptcmil: Multiple instance learning via prompt token clustering for whole slide image analysis. *Medical Image Analysis*, 2026.

Weiqin Zhao, Ziyu Guo, Yinshuang Fan, Yuming Jiang, Maximus CF Yeung, and Lequan Yu. Aligning knowledge concepts to whole slide images for precise histopathology image analysis. *npj Digital Medicine*, 7(1):383, 2024.

A Proofs and Evidence Identities

We first establish the validity of the morphology distributions, and then derive the class-balanced PMI interpretation, its KL-divergence-based decision equivalence, the cross-entropy geometry, and the exact tile-level score and class-margin decompositions.

A.1 CLASS-BALANCED EMPIRICAL PMI

Lemma 1 (Class-balanced empirical joint distribution and PMI)**.** *For each branch b, let the class prior be uniform, and define the class-conditional morphology distribution by*

$$P_b(y) = \frac{1}{C}, \qquad P_b(k \mid y) = \boldsymbol{q}_{yb}[k],$$

The pointwise mutual information of the joint outcome (y, k) is

$$\log \frac{P_b(y,k)}{P_b(y)P_b(k)} = \log \frac{\boldsymbol{q}_{yb}[k]}{\boldsymbol{m}_b[k]} = \boldsymbol{e}_{yb}[k]. \tag{19}$$

Hence, the diagnostic evidence in Equation 12 is exactly the pointwise mutual information of the class–morphology outcome (y, k) under the class-balanced empirical distribution P_b.

Proof. Since each class-conditional prototype $\boldsymbol{q}_{yb} \in \Delta^{K-1}$,

$$\sum_{k=1}^{K} \boldsymbol{q}_{yb}[k] = 1.$$

Therefore,

$$\sum_{y=1}^{C}\sum_{k=1}^{K} P_b(y,k) = \frac{1}{C}\sum_{y=1}^{C}\sum_{k=1}^{K} \boldsymbol{q}_{yb}[k] = \frac{1}{C}\sum_{y=1}^{C} 1 = 1, \tag{20}$$

so P_b is a valid probability mass function on $[C] \times [K]$.

Its class marginal is

$$\sum_{k=1}^{K} P_b(y,k) = \frac{1}{C}\sum_{k=1}^{K} \boldsymbol{q}_{yb}[k] = \frac{1}{C}. \tag{21}$$

Under the class-balanced prior, its morphology marginal can be written as

$$\sum_{y=1}^{C} P_b(y,k) = \frac{1}{C}\sum_{y=1}^{C} \boldsymbol{q}_{yb}[k] = \boldsymbol{m}_b[k]. \tag{22}$$

Substituting these quantities into the definition of pointwise mutual information gives

$$\log \frac{P_b(y,k)}{P_b(y)P_b(k)} = \log \frac{\frac{1}{C}\boldsymbol{q}_{yb}[k]}{\frac{1}{C}\boldsymbol{m}_b[k]} = \log \frac{\boldsymbol{q}_{yb}[k]}{\boldsymbol{m}_b[k]} = \boldsymbol{e}_{yb}[k], \tag{23}$$

which is exactly Equation 12. □

A.2 PMI–PROTOTYPE EQUIVALENCE

Proposition 1 (PMI–KL identity and predictive equivalence)**.** *For every slide X_s and candidate class y,*

$$s_y(X_s) = \frac{1}{B}\sum_{b=1}^{B} \left[\mathsf{KL}(\boldsymbol{p}_{sb} \,\|\, \boldsymbol{m}_b) - \mathsf{KL}(\boldsymbol{p}_{sb} \,\|\, \boldsymbol{q}_{yb})\right]. \tag{24}$$

Consequently,

$$\arg\max_{y\in[C]} s_y(X_s) = \arg\min_{y\in[C]} \frac{1}{B}\sum_{b=1}^{B} \mathsf{KL}(\boldsymbol{p}_{sb} \,\|\, \boldsymbol{q}_{yb}). \tag{25}$$

Proof. For a fixed branch b,

$$\begin{aligned}
&\mathsf{KL}(\boldsymbol{p}_{sb} \,\|\, \boldsymbol{m}_b) - \mathsf{KL}(\boldsymbol{p}_{sb} \,\|\, \boldsymbol{q}_{yb}) \\
&= \sum_{k=1}^{K} \boldsymbol{p}_{sb}[k] \left[\log \frac{\boldsymbol{p}_{sb}[k]}{\boldsymbol{m}_b[k]} - \log \frac{\boldsymbol{p}_{sb}[k]}{\boldsymbol{q}_{yb}[k]} \right] \\
&= \sum_{k=1}^{K} \boldsymbol{p}_{sb}[k] \log \frac{\boldsymbol{q}_{yb}[k]}{\boldsymbol{m}_b[k]} \\
&= \sum_{k=1}^{K} \boldsymbol{p}_{sb}[k] \boldsymbol{e}_{yb}[k]. \qquad (26)
\end{aligned}$$

Averaging Equation 26 over all branches and using Equation 13 gives Equation 24.

The first term in Equation 24,

$$\frac{1}{B} \sum_{b=1}^{B} \mathsf{KL}(\boldsymbol{p}_{sb} \,\|\, \boldsymbol{m}_b)\,,$$

is independent of the candidate class y. Therefore,

$$\arg\max_{y \in [C]} s_y(X_s) = \arg\min_{y \in [C]} \frac{1}{B} \sum_{b=1}^{B} \mathsf{KL}(\boldsymbol{p}_{sb} \,\|\, \boldsymbol{q}_{yb})\,,$$

which proves Equation 25. □

Corollary 1 (Cross-entropy in KL geometry)**.** *Define the class-y prototype distance as*

$$d_y(X_s) = \frac{1}{B} \sum_{b=1}^{B} \mathsf{KL}(\boldsymbol{p}_{sb} \,\|\, \boldsymbol{q}_{yb})\,, \qquad \boldsymbol{d}(X_s) = \big[d_1(X_s), \ldots, d_C(X_s)\big]^\top. \qquad (27)$$

Let

$$\boldsymbol{s}(X_s) = \big[s_1(X_s), \ldots, s_C(X_s)\big]^\top.$$

Then

$$\sigma(\boldsymbol{s}(X_s)) = \sigma(-\boldsymbol{d}(X_s))\,. \qquad (28)$$

For a labeled slide (X_s, y_s), let

$$\widehat{\boldsymbol{\eta}}_s = \sigma(\boldsymbol{s}(X_s)) = \sigma(-\boldsymbol{d}(X_s)) \in \Delta^{C-1}. \qquad (29)$$

The cross-entropy loss can then be written exactly as

$$\mathcal{L}_{\mathrm{CE}} = \log \left[1 + \sum_{c \neq y_s} \exp(d_{y_s}(X_s) - d_c(X_s)) \right]. \qquad (30)$$

Moreover, for every competitor $c \neq y_s$, the loss is strictly increasing in the distance gap

$$g_c(X_s) = d_{y_s}(X_s) - d_c(X_s) \qquad (31)$$

when the other gaps are fixed.

Proof. Equation 24 gives

$$s_y(X_s) = \frac{1}{B} \sum_{b=1}^{B} \mathsf{KL}(\boldsymbol{p}_{sb} \,\|\, \boldsymbol{m}_b) - d_y(X_s).$$

The first term is identical for all candidate classes. Since softmax is invariant to adding the same scalar to every coordinate,

$$\sigma(\boldsymbol{s}(X_s)) = \sigma(-\boldsymbol{d}(X_s))\,,$$

which proves Equation 28. Therefore,

$$\begin{aligned}\mathcal{L}_{\mathrm{CE}} &= -\log \widehat{\boldsymbol{\eta}}_s[y_s] \\ &= -\log \frac{\exp(-d_{y_s}(X_s))}{\sum_{c=1}^{C}\exp(-d_c(X_s))} \\ &= \log\left[\sum_{c=1}^{C}\exp(d_{y_s}(X_s)-d_c(X_s))\right] \\ &= \log\left[1+\sum_{c\neq y_s}\exp(d_{y_s}(X_s)-d_c(X_s))\right], \end{aligned} \tag{32}$$

which proves Equation 30. Using the distance gaps in Equation 31,

$$\mathcal{L}_{\mathrm{CE}} = \log\left[1+\sum_{r\neq y_s}\exp(g_r(X_s))\right]. \tag{33}$$

For any competitor $c \neq y_s$,

$$\frac{\partial \mathcal{L}_{\mathrm{CE}}}{\partial g_c(X_s)} = \frac{\exp(g_c(X_s))}{1+\sum_{r\neq y_s}\exp(g_r(X_s))} > 0. \tag{34}$$

Hence, minimizing cross-entropy decreases $d_{y_s}(X_s) - d_c(X_s)$, or equivalently increases the distance advantage of the ground-truth class over each competing prototype. □

A.3 Exact Tile-Level Score and Class-Margin Decomposition

Proposition 2 (Exact tile-level score and class-margin decomposition)**.** *For class $y \in [C]$, define the class-y contribution of tile i by*

$$\tau_{si}^{y} = \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{\alpha}_{sb}[i]\sum_{k=1}^{K}\boldsymbol{\psi}_{si}[k]\boldsymbol{e}_{yb}[k]. \tag{35}$$

Then

$$s_y(X_s) = \sum_{i=1}^{N_s}\tau_{si}^{y}. \tag{36}$$

For any two classes $c, r \in [C]$, the corresponding tile-level pairwise contribution satisfies

$$\tau_{si}^{c} - \tau_{si}^{r} = \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{\alpha}_{sb}[i]\sum_{k=1}^{K}\boldsymbol{\psi}_{si}[k]\log\frac{\boldsymbol{q}_{cb}[k]}{\boldsymbol{q}_{rb}[k]}. \tag{37}$$

Consequently,

$$s_c(X_s) - s_r(X_s) = \sum_{i=1}^{N_s}\left(\tau_{si}^{c} - \tau_{si}^{r}\right). \tag{38}$$

If $\widehat{\boldsymbol{\eta}}_s = \sigma(\boldsymbol{s}(X_s))$, then the predictive log-odds satisfy

$$\log\frac{\widehat{\boldsymbol{\eta}}_s[c]}{\widehat{\boldsymbol{\eta}}_s[r]} = s_c(X_s) - s_r(X_s) = \sum_{i=1}^{N_s}\left(\tau_{si}^{c} - \tau_{si}^{r}\right). \tag{39}$$

Proof. Substituting Equation 9 into Equation 13 and exchanging the finite sums gives

$$\begin{aligned} s_y(X_s) &= \frac{1}{B}\sum_{b=1}^{B}\sum_{k=1}^{K}\left(\sum_{i=1}^{N_s}\boldsymbol{\alpha}_{sb}[i]\boldsymbol{\psi}_{si}[k]\right)\boldsymbol{e}_{yb}[k] \\ &= \sum_{i=1}^{N_s}\frac{1}{B}\sum_{b=1}^{B}\boldsymbol{\alpha}_{sb}[i]\sum_{k=1}^{K}\boldsymbol{\psi}_{si}[k]\boldsymbol{e}_{yb}[k] \\ &= \sum_{i=1}^{N_s}\tau_{si}^{y}, \end{aligned} \tag{40}$$

which proves Equation 36. For two classes c and r, the class-agnostic marginal cancels from the difference between their diagnostic evidence:

$$\begin{aligned} \boldsymbol{e}_{cb}[k] - \boldsymbol{e}_{rb}[k] &= \log\frac{\boldsymbol{q}_{cb}[k]}{\boldsymbol{m}_b[k]} - \log\frac{\boldsymbol{q}_{rb}[k]}{\boldsymbol{m}_b[k]} \\ &= \log\frac{\boldsymbol{q}_{cb}[k]}{\boldsymbol{q}_{rb}[k]}. \end{aligned} \tag{41}$$

Therefore,

$$\begin{aligned} \tau_{si}^{c} - \tau_{si}^{r} &= \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{\alpha}_{sb}[i]\sum_{k=1}^{K}\boldsymbol{\psi}_{si}[k]\left(\boldsymbol{e}_{cb}[k] - \boldsymbol{e}_{rb}[k]\right) \\ &= \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{\alpha}_{sb}[i]\sum_{k=1}^{K}\boldsymbol{\psi}_{si}[k]\log\frac{\boldsymbol{q}_{cb}[k]}{\boldsymbol{q}_{rb}[k]}, \end{aligned} \tag{42}$$

which proves Equation 37. Using Equation 36,

$$\begin{aligned} s_c(X_s) - s_r(X_s) &= \sum_{i=1}^{N_s}\tau_{si}^{c} - \sum_{i=1}^{N_s}\tau_{si}^{r} \\ &= \sum_{i=1}^{N_s}\left(\tau_{si}^{c} - \tau_{si}^{r}\right), \end{aligned} \tag{43}$$

which proves Equation 38. Finally, since $\widehat{\boldsymbol{\eta}}_s = \sigma(\boldsymbol{s}(X_s))$,

$$\log\frac{\widehat{\boldsymbol{\eta}}_s[c]}{\widehat{\boldsymbol{\eta}}_s[r]} = \log\frac{\exp(s_c(X_s))}{\exp(s_r(X_s))} = s_c(X_s) - s_r(X_s), \tag{44}$$

which, together with Equation 38, proves Equation 39. □

Equations 36 and 39 are exact algebraic decompositions of the model's class score and predictive log-odds, respectively. They establish the computational faithfulness of MELD's tile-level contributions.

B Training and Numerical Implementation

B.1 Datasets.

We evaluate all datasets as single-label WSI classification tasks. CAMELYON16 is formulated as binary classification between normal and tumor slides. CAMELYON17 is formulated as a four-class slide-level classification task among normal, isolated tumor cell (ITC), micro-metastasis, and macro-metastasis cases. For BRACS, we group the diagnostic labels into three classes: benign, atypical, and malignant. For CPTAC, we perform binary lung-cancer subtype classification between lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LSCC). The data partitioning protocol is specified separately for each dataset. For CAMELYON16, we retain the official test set and split the official training set into training and validation subsets with a 9:1 ratio. For CAMELYON17,

we use Centers 0–2 as the development cohort and Centers 3–4 as a fixed held-out test cohort. The patients from Centers 0–2 are further divided into training and validation sets with an 8:2 ratio. Because each CAMELYON17 patient can contribute multiple WSIs, this split is performed at the patient level, so that all WSIs from the same patient remain in the same partition. For BRACS, we use the official training, validation, and test partitions provided with the dataset; these partitions are patient-disjoint. For CPTAC, which does not provide an official classification split, we construct an 8:1:1 training/validation/test split at the patient level, with all WSIs from the same patient assigned to a single partition. Thus, no patient appears in more than one partition for either CAMELYON17 or CPTAC. The resulting splits are fixed and shared by MELD and all baseline methods. The validation set is used exclusively for checkpoint selection and hyperparameter selection, and the test set is not used for prototype estimation, checkpoint selection, or any other model-selection procedure. The exact patient/slide partitions used in all experiments are provided with the released code.

B.2 Prototype initialization.

The morphology codewords are initialized by K-means using sampled training-set tile representations only. Given the current network parameters, the branch-specific class prototypes $\boldsymbol{q}_{yb}$ are estimated from all training slides according to Equation 10. The corresponding morphology marginal $\boldsymbol{m}_b$ and PMI evidence $\boldsymbol{e}_{yb}$ are then computed directly according to Equations 11 and 12, respectively.

B.3 Epoch-wise prototype update.

We update the prototype statistics once per training epoch. Before the first epoch, a full pass over the training set using all B branches is used to initialize $\boldsymbol{q}_{yb}$, $\boldsymbol{m}_b$, and $\boldsymbol{e}_{yb}$. After each subsequent epoch, the updated model is evaluated over the complete training set using all B branches, and these statistics are recomputed. The refreshed prototype, marginal, and PMI tables are detached from the computational graph and kept fixed throughout the next epoch. Thus, gradients are never propagated through the empirical prototype estimation procedure.

B.4 Training objective.

MELD is trained with class-balanced slide sampling. For each sampled training slide (X_s, y_s), the class scores are computed using all B attention branches according to Equation 13. With the current prototype and PMI tables treated as stop-gradient statistics, the model is optimized using

$$\mathcal{L}_s = \mathcal{L}_{\mathrm{CE}}\big(\boldsymbol{s}(X_s), y_s\big) + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}}^{(s)}. \tag{45}$$

Here, $\mathcal{L}_{\mathrm{div}}^{(s)}$ is defined in Equation 6. Gradients are propagated through the feature reducer, attention branches, tokenizer, and morphology codewords, while the prototype and PMI statistics remain fixed within each epoch.

B.5 Optimization.

All models are trained for 100 epochs with a batch size of one. We use the Adam optimizer with $\eta = 10^{-3}$, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon_{\mathrm{Adam}} = 10^{-8}$, and zero weight decay. The learning rate is decayed using cosine annealing over the full training schedule.

B.6 Validation, checkpoint selection, and test evaluation.

After each training epoch, the prototype and PMI statistics are recomputed using the complete training set, and the resulting model is evaluated on the validation set using Equation 13. Class probabilities are obtained by applying softmax to the evidence scores. The checkpoint selected for final evaluation is the epoch achieving the largest sum of validation Macro-$F1$ and Macro-AUROC,

$$t^\star = \arg\max \left[\mathrm{MacroF1}_{\mathrm{val}}^{(t)} + \mathrm{MacroAUROC}_{\mathrm{val}}^{(t)}\right]. \tag{46}$$

Table 2: Paired bootstrap evaluation of the performance difference between MELD and the strongest baseline for each dataset–encoder setting. For each bootstrap replicate, the same resampled test cases are used for both methods and all five training seeds. The paired differences are averaged across seeds within each replicate, and the 2.5th and 97.5th percentiles across bootstrap replicates define the 95% confidence interval.

Feature	$F1$ Runner-up	Macro-$F1$ Δ [95% CI]	AUROC Runner-up	Macro-AUROC Δ [95% CI]
BRACS ImageNet	MHIM-MIL	-0.001 $[-0.070,\ 0.068]$	MHIM-MIL	-0.024 $[-0.062,\ 0.014]$
BRACS medical	CAMIL	0.036 $[0.002,\ 0.076]$	MHIM-MIL	0.007 $[0.002,\ 0.010]$
CAMELYON16 ImageNet	DTFD-MIL	0.007 $[0.004,\ 0.023]$	PNEA-MIL	0.010 $[0.002,\ 0.043]$
CAMELYON16 medical	DTFD-MIL	0.009 $[0.003,\ 0.025]$	Additive MIL	0.003 $[0.001,\ 0.009]$
CAMELYON17 ImageNet	ASMIL	0.015 $[0.003,\ 0.039]$	ASMIL	0.028 $[0.001,\ 0.059]$
CAMELYON17 medical	CAMIL	-0.010 $[-0.111,\ 0.091]$	ACMIL	-0.010 $[-0.057,\ 0.037]$
CPTAC ImageNet	PNEA-MIL	0.007 $[0.001,\ 0.021]$	PNEA-MIL	-0.008 $[-0.019,\ 0.003]$
CPTAC medical	PNEA-MIL	0.018 $[0.007,\ 0.029]$	PNEA-MIL	0.001 $[0.001,\ 0.003]$

After training is complete, the parameters from epoch $t^\star$ are restored, the corresponding prototype and PMI statistics are recomputed using the training set only, and the resulting checkpoint is evaluated once on the test set. The test set is therefore not used for prototype estimation, checkpoint selection, or any other model-selection decision.

B.7 Statistical evaluation.

In addition to reporting the mean and standard deviation across the five random seeds, we quantify the uncertainty of the performance differences using paired bootstrap resampling of the test set. For each dataset–encoder setting and evaluation metric, MELD is compared with the strongest baseline in Table 1. For each bootstrap replicate $r \in \{1, \dots, 5000\}$, we draw one stratified bootstrap sample of the test set and evaluate both MELD and the baseline on exactly the same resampled cases for all five random seeds. For seed s, we compute the paired performance difference

$$\Delta_{\mathcal{M}}^{(r,s)} = \mathcal{M}_{\mathrm{MELD}}^{(r,s)} - \mathcal{M}_{\mathrm{Baseline}}^{(r,s)}, \tag{47}$$

where $\mathcal{M}$ denotes either Macro-$F1$ or Macro-AUROC. We then average the paired differences across the five independently trained seeds within each bootstrap replicate,

$$\bar{\Delta}_{\mathcal{M}}^{(r)} = \frac{1}{5} \sum_{s=1}^{5} \Delta_{\mathcal{M}}^{(r,s)}. \tag{48}$$

The 2.5th and 97.5th percentiles of $\{\bar{\Delta}_{\mathcal{M}}^{(r)}\}_{r=1}^{5000}$ are reported as the 95% confidence interval. Table 2 reports the resulting confidence intervals. For Macro-$F1$, the 95% CI lies entirely above zero in six of the eight dataset–encoder settings, BRACS with medical SSL features, CAMELYON16 with both encoders, CAMELYON17 with ImageNet features, and CPTAC with both encoders. The intervals overlap zero for BRACS with ImageNet features and CAMELYON17 with medical SSL features. For Macro-AUROC, the 95% CI lies entirely above zero in five settings: BRACS with medical SSL features, CAMELYON16 with both encoders, CAMELYON17 with ImageNet features, and CPTAC with medical SSL features. The intervals overlap zero for BRACS with ImageNet features, CAMELYON17 with medical SSL features, and CPTAC with ImageNet features. Overall, the bootstrap analysis provides evidence that the Macro-$F1$ improvements are more consistent across dataset–encoder settings than the Macro-AUROC improvements, while also identifying settings in which the performance difference is not clearly distinguishable from zero under test-set resampling.

C Macro-AUROC and Macro-$F1$ under Class Imbalance

Since the datasets used in this work are class-imbalanced, we report Macro-$F1$ and Macro-AUROC as the primary evaluation metrics. Both metrics assign equal weight to each diagnostic class, thereby reducing the influence of class-frequency imbalance on the overall performance.

Let $\mathcal{Y} = \{1, \dots, C\}$ denote the set of diagnostic classes. For a sample x with ground-truth label y, let $\boldsymbol{s}(x) \in \mathbb{R}^C$ denote the model scores and let

$$\boldsymbol{p}(x) = \sigma(\boldsymbol{s}(x)) \tag{49}$$

denote the corresponding class probabilities. We use TorchMetrics to compute both metrics.

C.1 MACRO-$F1$

All datasets are treated as single-label classification tasks. The predicted class is obtained as

$$\hat{y} = \arg\max_{c \in [C]} \boldsymbol{p}(x)[c] = \arg\max_{c \in [C]} \boldsymbol{s}(x)[c]. \tag{50}$$

For each class c, we compute precision and recall in a one-vs-rest manner:

$$\mathrm{Precision}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \qquad \mathrm{Recall}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}. \tag{51}$$

The class-wise $F1$ score is

$$F1_c = \frac{2\,\mathrm{Precision}_c\,\mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}, \tag{52}$$

and Macro-$F1$ is the uniform average over all classes:

$$\text{Macro-}F1 = \frac{1}{C}\sum_{c=1}^{C} F1_c. \tag{53}$$

Thus, each WSI is assigned exactly one predicted class using the $\arg\max$ decision rule.

C.2 MACRO-AUROC

For each class c, we construct a one-vs-rest ROC curve using $\boldsymbol{p}(x)[c]$ as the continuous prediction score and $\mathbf{1}[y = c]$ as the binary target, where $\mathbf{1}[\cdot]$ denotes the indicator function. Let AUROC_c denote the resulting area under the ROC curve. Macro-AUROC is defined as

$$\text{Macro-AUROC} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{AUROC}_c. \tag{54}$$

Unlike Macro-$F1$, Macro-AUROC is threshold-independent and measures the class-wise ranking quality of the predicted scores.

D FAITHFULNESS OF TILE-LEVEL DECISION CONTRIBUTIONS

The exact score and class-margin decompositions derived in Appendix A.3 establish the computational faithfulness of MELD's tile-level decision contributions, which exactly reconstruct the model score and pairwise decision margin. We further examine whether these contributions are behaviorally faithful. Specifically, if a tile receives a large positive contribution to a decision, removing that tile should weaken the corresponding decision more strongly than removing tiles selected by attention alone or at random. For each test slide X_s, we first evaluate MELD on the complete bag and fix the predicted class

$$c = \arg\max_{y} s_y(X_s) \tag{55}$$

and its strongest competing class

$$r = \arg\max_{y \neq c} s_y(X_s). \tag{56}$$

We then compute the contribution of each tile i to the pairwise margin between c and r. Following Appendix A.3,

$$\tau_{si}^c - \tau_{si}^r = \frac{1}{B} \sum_{b=1}^{B} \boldsymbol{\alpha}_{sb}[i] \sum_{k=1}^{K} \boldsymbol{\psi}_{si}[k] \left(\boldsymbol{e}_{cb}[k] - \boldsymbol{e}_{rb}[k] \right), \tag{57}$$

where $\alpha_{sb}[i]$ is the attention weight assigned to tile i by branch b, $\psi_{si}[k]$ is its soft assignment to morphology token k, and $e_{yb}[k]$ is the corresponding class-specific signed evidence. By construction,

$$s_c(X_s) - s_r(X_s) = \sum_{i=1}^{N_s} (\tau_{si}^c - \tau_{si}^r). \tag{58}$$

Tiles are ranked once in descending order of $\tau_{si}^c - \tau_{si}^r$. We progressively remove the top $1\%, 5\%, 10\%, 20\%, 30\%, 40\%$, and 50% of tiles and re-evaluate the complete model on each reduced bag. After every intervention, the full forward pass, including attention computation and normalization, is recomputed; the perturbed prediction is therefore not approximated by subtracting the original tile contributions. The original classes c and r are kept fixed throughout the experiment so that all deletion levels measure changes to the same initial decision.

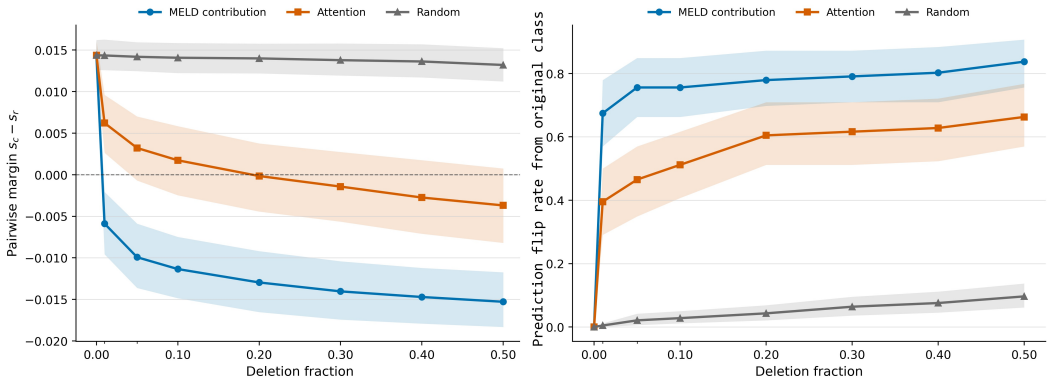

Figure 6: Behavioral faithfulness of MELD tile-level evidence on **BRACS** using **ResNet-18 ImageNet-pretrained features**. Tiles are progressively removed according to their MELD pairwise decision contribution, aggregated attention weight, or a random ranking, and the complete model is re-evaluated on each reduced bag. **Left:** remaining pairwise decision margin between the original predicted class and its strongest competitor. **Right:** fraction of slides whose prediction changes from the original predicted class. MELD-ranked deletion consistently produces a faster reduction in the original decision margin and more frequent prediction flips than attention-based or random deletion, showing that the decomposed tile-level contributions identify evidence that is more strongly coupled to the model's actual decision. Shaded regions denote 95% bootstrap confidence intervals.

We compare MELD ranking with two controls. Under *attention deletion*, tiles are ranked by their aggregated attention weight, while under *random deletion*, the same number of tiles is removed uniformly at random and the results are averaged over multiple random draws. The primary endpoint is the remaining pairwise decision margin

$$M_s(q) = s_c\left(X_s^{(-S_q)}\right) - s_r\left(X_s^{(-S_q)}\right), \tag{59}$$

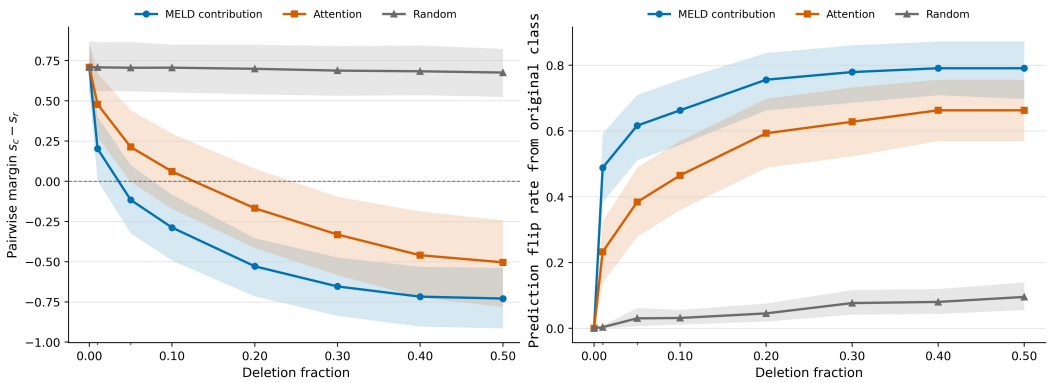

Figure 7: Behavioral faithfulness of MELD tile-level evidence on **BRACS** using **ViT-S SSL-pretrained features**. Tiles are progressively removed according to their MELD pairwise decision contribution, aggregated attention weight, or a random ranking, and the complete model is re-evaluated on each reduced bag. **Left:** remaining pairwise decision margin between the original predicted class and its strongest competitor. **Right:** fraction of slides whose prediction changes from the original predicted class. MELD-ranked deletion consistently produces a faster reduction in the original decision margin and more frequent prediction flips than attention-based or random deletion, showing that the decomposed tile-level contributions identify evidence that is more strongly coupled to the model's actual decision. Shaded regions denote 95% bootstrap confidence intervals.

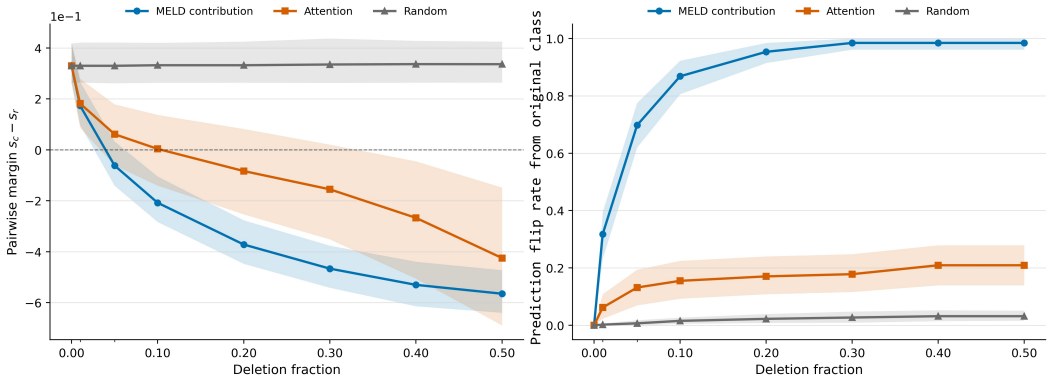

Figure 8: Behavioral faithfulness of MELD tile-level evidence on **CAMELYON16** using **ResNet-18 ImageNet-pretrained features**. Tiles are progressively removed according to their MELD pairwise decision contribution, aggregated attention weight, or a random ranking, and the complete model is re-evaluated on each reduced bag. **Left:** remaining pairwise decision margin between the original predicted class and its strongest competitor. **Right:** fraction of slides whose prediction changes from the original predicted class. MELD-ranked deletion consistently produces a faster reduction in the original decision margin and more frequent prediction flips than attention-based or random deletion, showing that the decomposed tile-level contributions identify evidence that is more strongly coupled to the model's actual decision. Shaded regions denote 95% bootstrap confidence intervals.

where S_q denotes the set of tiles removed at deletion fraction q. We additionally report the prediction-flip rate, defined as the fraction of slides whose predicted class changes from the original prediction c after deletion.

The deletion curves show a consistent pattern across all eight dataset–feature settings. As progressively more tiles are removed, the average pairwise margin decreases monotonically under all three deletion policies. However, the decrease is markedly faster when tiles are ranked by MELD's

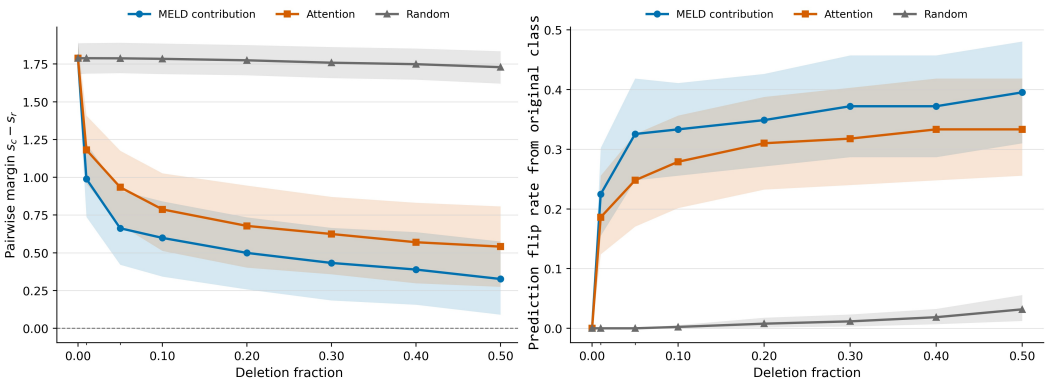

Figure 9: Behavioral faithfulness of MELD tile-level evidence on **CAMELYON16** using **ViT-S SSL-pretrained features**. Tiles are progressively removed according to their MELD pairwise decision contribution, aggregated attention weight, or a random ranking, and the complete model is re-evaluated on each reduced bag. **Left:** remaining pairwise decision margin between the original predicted class and its strongest competitor. **Right:** fraction of slides whose prediction changes from the original predicted class. MELD-ranked deletion consistently produces a faster reduction in the original decision margin and more frequent prediction flips than attention-based or random deletion, showing that the decomposed tile-level contributions identify evidence that is more strongly coupled to the model's actual decision. Shaded regions denote 95% bootstrap confidence intervals.

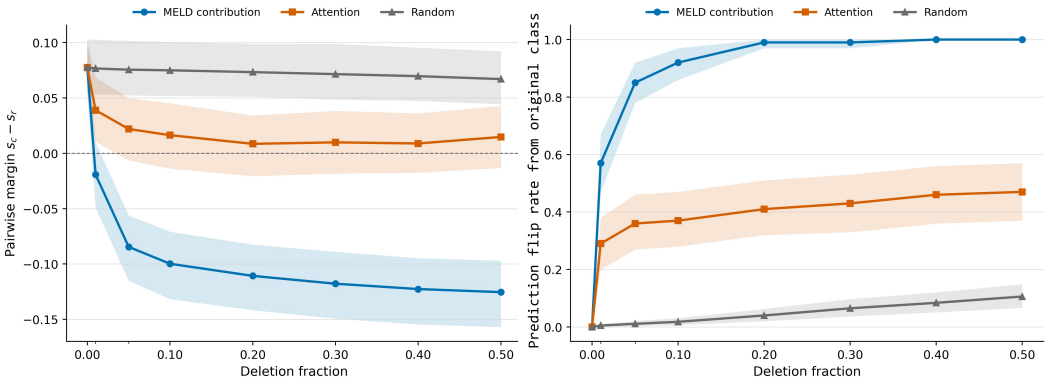

Figure 10: Behavioral faithfulness of MELD tile-level evidence on **CAMELYON17** using **ResNet-18 ImageNet-pretrained features**. Tiles are progressively removed according to their MELD pairwise decision contribution, aggregated attention weight, or a random ranking, and the complete model is re-evaluated on each reduced bag. **Left:** remaining pairwise decision margin between the original predicted class and its strongest competitor. **Right:** fraction of slides whose prediction changes from the original predicted class. MELD-ranked deletion consistently produces a faster reduction in the original decision margin and more frequent prediction flips than attention-based or random deletion, showing that the decomposed tile-level contributions identify evidence that is more strongly coupled to the model's actual decision. Shaded regions denote 95% bootstrap confidence intervals.

pairwise evidence contribution. Attention-based deletion produces a substantially slower reduction, while random deletion typically has the smallest effect. This separation is already visible at small deletion fractions and becomes increasingly pronounced as more tiles are removed, indicating that MELD concentrates the class-discriminative evidence into tiles whose removal directly suppresses the original decision rather than merely identifying highly attended regions.

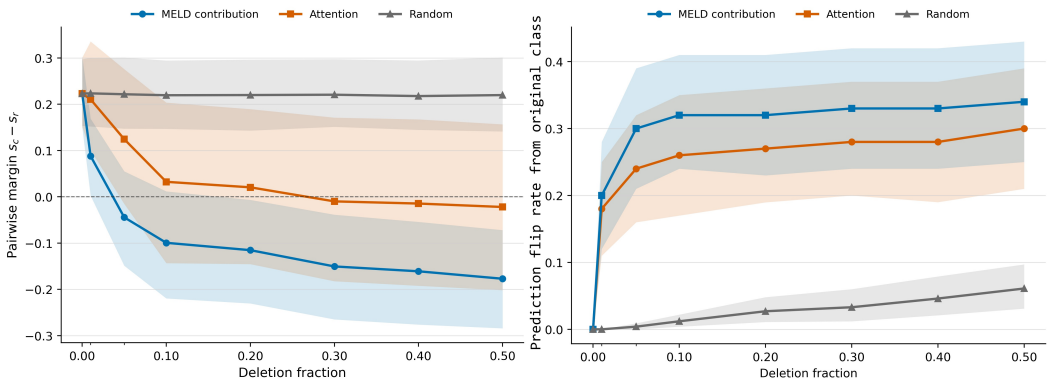

Figure 11: Behavioral faithfulness of MELD tile-level evidence on **CAMELYON17** using **ViT-S SSL-pretrained features**. Tiles are progressively removed according to their MELD pairwise decision contribution, aggregated attention weight, or a random ranking, and the complete model is re-evaluated on each reduced bag. **Left:** remaining pairwise decision margin between the original predicted class and its strongest competitor. **Right:** fraction of slides whose prediction changes from the original predicted class. MELD-ranked deletion consistently produces a faster reduction in the original decision margin and more frequent prediction flips than attention-based or random deletion, showing that the decomposed tile-level contributions identify evidence that is more strongly coupled to the model's actual decision. Shaded regions denote 95% bootstrap confidence intervals.

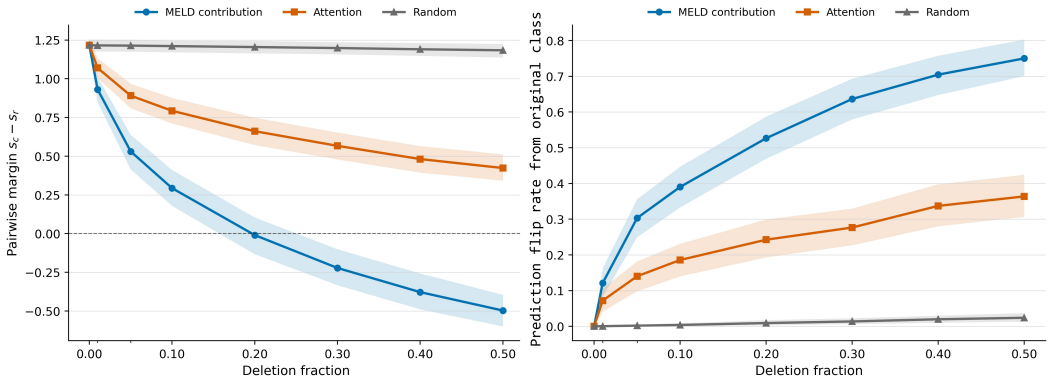

Figure 12: Behavioral faithfulness of MELD tile-level evidence on **CPTAC** using **ResNet-18 ImageNet-pretrained features**. Tiles are progressively removed according to their MELD pairwise decision contribution, aggregated attention weight, or a random ranking, and the complete model is re-evaluated on each reduced bag. **Left:** remaining pairwise decision margin between the original predicted class and its strongest competitor. **Right:** fraction of slides whose prediction changes from the original predicted class. MELD-ranked deletion consistently produces a faster reduction in the original decision margin and more frequent prediction flips than attention-based or random deletion, showing that the decomposed tile-level contributions identify evidence that is more strongly coupled to the model's actual decision. Shaded regions denote 95% bootstrap confidence intervals.

The prediction-flip results show the same behavior from a complementary perspective. As the deletion fraction increases, substantially more slides change their prediction away from the original class under contribution-ranked deletion than under either attention-based or random deletion. In the representative setting shown in Fig. 13, removing only 10% of the highest-contribution tiles changes the prediction for approximately 18% of test slides, compared with approximately 6% under attention deletion and almost no changes under random deletion. At 50% deletion, the flip rate reaches approximately 45% for MELD ranking, while remaining around 11% for attention ranking and close to

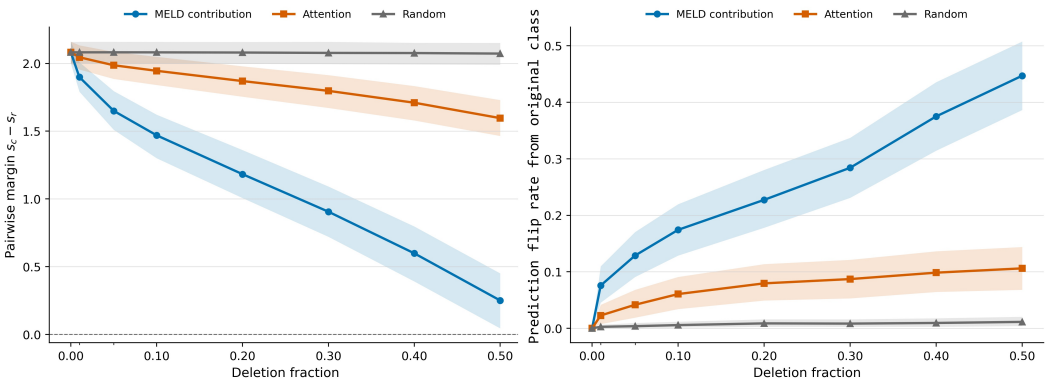

Figure 13: Behavioral faithfulness of MELD tile-level evidence on **CPTAC** using **ViT-S SSL-pretrained features**. Tiles are progressively removed according to their MELD pairwise decision contribution, aggregated attention weight, or a random ranking, and the complete model is re-evaluated on each reduced bag. **Left:** remaining pairwise decision margin between the original predicted class and its strongest competitor. **Right:** fraction of slides whose prediction changes from the original predicted class. MELD-ranked deletion consistently produces a faster reduction in the original decision margin and more frequent prediction flips than attention-based or random deletion, showing that the decomposed tile-level contributions identify evidence that is more strongly coupled to the model's actual decision. Shaded regions denote 95% bootstrap confidence intervals.

zero for random removal. Consistently, the average pairwise margin decreases from approximately 2.1 on the complete bags to about 0.25 after removing 50% of MELD-ranked tiles, compared with approximately 1.6 under attention deletion and only a small reduction under random deletion.

We additionally observe a consistent difference between the two feature encoders. Before any deletion ($q = 0$), the initial pairwise margin between the predicted class and its strongest competitor is higher with medical SSL features than with ImageNet-pretrained features across all four datasets. This indicates stronger separation between the selected class and its nearest competitor in MELD's decision space when pathology-specific representations are used. The observation is qualitatively consistent with the classification results in Table 1, where medical SSL features also yield higher Macro-F1 and Macro-AUROC than ImageNet-pretrained features across the four benchmarks. Because margins from separately trained models are not necessarily calibrated to a common numerical scale, we interpret this agreement as a consistent empirical trend rather than a direct quantitative relationship between margin magnitude and classification performance. Together with the exact decomposition in Appendix A.3, these experiments provide two complementary forms of validation. MELD's tile-level evidence is *computationally faithful* because the contributions exactly compose the model's score and pairwise margin, and *behaviorally faithful* because removing tiles with large decision contributions causes the decision margin to collapse substantially faster and the original prediction to flip substantially more frequently than under attention-based or random interventions.

E Ablation Study

We complement the component-level evaluation in the main text with controlled sensitivity analyses of the principal design hyperparameters. For each sensitivity analysis below, we vary only the hyperparameter under study while holding all other hyperparameters fixed to the configuration used in the main experiments, and evaluate all variants using the same five-seed protocol. We report Macro-$F1$ and Macro-AUROC as mean $\pm$ standard deviation over five seeds, consistent with the primary evaluation protocol. These experiments address the robustness of the morphology vocabulary, soft token assignment, and tokenizer capacity.

Table 3: Component ablation of MELD on CPTAC using ImageNet-pretrained features. All variants are evaluated under the same training and evaluation protocol. **Bold** and underlined values denote the best and second-best results, respectively.

Variant	Attention	Tokenizer	PMI classifier	Macro-$F1\uparrow$	Macro-AUROC $\uparrow$
Full MELD	✓	✓	✓	$\mathbf{0.914}_{\pm 0.011}$	$\mathbf{0.965}_{\pm 0.009}$
w/o morphology tokenizer	✓	×	×	$\underline{0.901}_{\pm 0.004}$	$0.955_{\pm 0.003}$
Token distribution + linear classifier	✓	✓	×	$0.890_{\pm 0.012}$	$\underline{0.960}_{\pm 0.003}$
Uniform tile weighting	×	✓	✓	$0.877_{\pm 0.014}$	$0.942_{\pm 0.002}$
Single attention branch ($B = 1$)	✓	✓	✓	$0.886_{\pm 0.014}$	$0.949_{\pm 0.007}$
w/o diversity loss ($\lambda_{\text{div}} = 0$)	✓	✓	✓	$0.901_{\pm 0.017}$	$0.957_{\pm 0.010}$

Table 4: Component ablation of MELD on CPTAC using medical SSL features. All variants are evaluated under the same training and evaluation protocol. **Bold** and underlined values denote the best and second-best results, respectively.

Variant	Attention	Tokenizer	PMI classifier	Macro-$F1\uparrow$	Macro-AUROC $\uparrow$
Full MELD	✓	✓	✓	$\mathbf{0.945}_{\pm 0.008}$	$\mathbf{0.983}_{\pm 0.001}$
w/o morphology tokenizer	✓	×	×	$0.936_{\pm 0.010}$	$0.980_{\pm 0.002}$
Token distribution + linear classifier	✓	✓	×	$\underline{0.938}_{\pm 0.013}$	$\underline{0.980}_{\pm 0.000}$
Uniform tile weighting	×	✓	✓	$0.887_{\pm 0.007}$	$0.959_{\pm 0.001}$
Single attention branch ($B = 1$)	✓	✓	✓	$0.930_{\pm 0.011}$	$0.975_{\pm 0.006}$
w/o diversity loss ($\lambda_{\text{div}} = 0$)	✓	✓	✓	$0.931_{\pm 0.013}$	$0.977_{\pm 0.005}$

E.1 Component ablation.

We first examine the contribution of the principal design mechanisms in MELD. Starting from the full model, we remove or replace one design mechanism at a time while preserving the remaining applicable components and keeping the training and evaluation protocol unchanged. In particular, we examine the morphology tokenizer, the PMI-based evidence classifier, attention-based evidence collection, multiple attention branches, and the attention-diversity regularizer. Because the PMI classifier operates on the morphology-token distribution produced by the tokenizer, removing the morphology tokenizer necessarily eliminates the PMI-based decision rule. In contrast, we compare the empirical prototype-based decision rule with a learned parametric classifier by retaining the morphology-token distribution and replacing the proposed decision rule with a linear classification head. Tables 3 and 4 report the results on CPTAC, while Tables 5 and 6 report the corresponding results on BRACS.

The full MELD model achieves the strongest performance for both datasets and both feature representations. On CPTAC, MELD obtains Macro-$F1$/Macro-AUROC of $0.914/0.965$ with ImageNet-pretrained features and $0.945/0.983$ with medical SSL features. The same trend holds on the more challenging BRACS benchmark, where the full model achieves $0.529/0.764$ and $0.709/0.878$, respectively. Removing the morphology tokenizer together with the PMI-based decision rule consistently reduces performance, yielding $0.901/0.955$ and $0.936/0.980$ on CPTAC, and $0.518/0.726$ and $0.634/0.851$ on BRACS. These results show that explicitly representing and classifying slides through morphology-token distributions provides a consistent benefit over the corresponding attention-based model without this distributional evidence representation.

To compare the empirical prototype-based decision rule with a learned parametric classifier, we retain the morphology-token distribution and replace the proposed decision rule with a linear classification head. This variant remains consistently below the full MELD model. Its Macro-$F1$/Macro-AUROC is $0.890/0.960$ and $0.938/0.980$ on CPTAC, and $0.509/0.704$ and $0.671/0.866$ on BRACS. The difference is particularly pronounced on BRACS with ImageNet features, where replacing PMI with a linear classifier decreases Macro-AUROC from 0.764 to 0.704. Thus, the benefit of the morphology representation is not explained simply by introducing a discrete intermediate representation and training an additional parametric classifier. Instead, the results support the empirical prototype-based decision rule over a learned linear classifier on the same morphology-token representation.

The evidence-collection mechanism is also consistently important. Replacing learned attention with uniform tile weighting produces the largest degradation among the examined variants in ev-

Table 5: Component ablation of MELD on BRACS using ImageNet-pretrained features. All variants are evaluated under the same training and evaluation protocol. **Bold** and underlined values denote the best and second-best results, respectively.

Variant	Attention	Tokenizer	PMI classifier	Macro-$F1$ ↑	Macro-AUROC ↑
Full MELD	✓	✓	✓	$\mathbf{0.529}_{\pm 0.043}$	$\mathbf{0.764}_{\pm 0.023}$
w/o morphology tokenizer	✓	×	×	$\underline{0.518}_{\pm 0.044}$	$0.726_{\pm 0.013}$
Token distribution + linear classifier	✓	✓	×	$0.509_{\pm 0.052}$	$0.704_{\pm 0.036}$
Uniform tile weighting	×	✓	✓	$0.424_{\pm 0.057}$	$0.671_{\pm 0.041}$
Single attention branch ($B = 1$)	✓	✓	✓	$0.509_{\pm 0.024}$	$0.701_{\pm 0.010}$
w/o diversity loss ($\lambda_{\text{div}} = 0$)	✓	✓	✓	$0.514_{\pm 0.012}$	$\underline{0.753}_{\pm 0.013}$

Table 6: Component ablation of MELD on BRACS using medical SSL features. All variants are evaluated under the same training and evaluation protocol. **Bold** and underlined values denote the best and second-best results, respectively.

Variant	Attention	Tokenizer	PMI classifier	Macro-$F1$ ↑	Macro-AUROC ↑
Full MELD	✓	✓	✓	$\mathbf{0.709}_{\pm 0.020}$	$\mathbf{0.878}_{\pm 0.011}$
w/o morphology tokenizer	✓	×	×	$0.634_{\pm 0.035}$	$0.851_{\pm 0.012}$
Token distribution + linear classifier	✓	✓	×	$\underline{0.671}_{\pm 0.032}$	$\underline{0.866}_{\pm 0.009}$
Uniform tile weighting	×	✓	✓	$0.579_{\pm 0.013}$	$0.820_{\pm 0.002}$
Single attention branch ($B = 1$)	✓	✓	✓	$0.654_{\pm 0.051}$	$0.872_{\pm 0.014}$
w/o diversity loss ($\lambda_{\text{div}} = 0$)	✓	✓	✓	$0.692_{\pm 0.004}$	$0.870_{\pm 0.017}$

ery dataset–encoder setting. On CPTAC, Macro-$F1$ decreases from 0.914 to 0.877 with ImageNet features and from 0.945 to 0.887 with medical SSL features, while Macro-AUROC decreases from 0.965 to 0.942 and from 0.983 to 0.959, respectively. The effect is even larger on BRACS: Macro-$F1$ decreases from 0.529 to 0.424 and from 0.709 to 0.579, while Macro-AUROC decreases from 0.764 to 0.671 and from 0.878 to 0.820. These results support selectively concentrating the morphology distribution on diagnostically informative regions rather than treating all tissue tiles equally.

Multi-branch evidence collection provides an additional benefit. Restricting MELD to a single attention branch consistently reduces both metrics across the four settings, including Macro-$F1$ reductions from 0.914 to 0.886 on CPTAC with ImageNet features and from 0.709 to 0.654 on BRACS with medical SSL features. For this single-branch variant, the diversity regularizer is set to zero because no pairwise branch comparison exists. Removing the diversity regularizer from the multi-branch model also consistently decreases performance, with the full model outperforming the unregularized variant on both metrics for both datasets and feature representations. Together, these results indicate that multiple attention branches capture complementary evidence and that explicitly discouraging redundant branch distributions improves their joint representation.

Overall, the component-wise ablations consistently support the three-module design of MELD: learned attention determines *where* evidence is collected, the morphology tokenizer determines *what* morphology is represented, and the PMI-based classifier determines *how* the observed morphology supports or opposes each diagnosis. No individual mechanism alone accounts for the performance of the full model. Instead, the consistent degradation caused by uniform evidence collection, using a single attention branch, removal of diversity regularization, removal of the morphology-based representation, or replacement of the PMI decision rule indicates that MELD benefits from their joint integration within a common distributional evidence framework.

E.2 Morphology vocabulary size.

The number of morphology codewords K controls the granularity of the morphology-token vocabulary. We evaluate $K \in \{4, 8, 16, 24, 32, 48, 64\}$ on BRACS and CPTAC using both ImageNet-pretrained ResNet-18 and self-supervised ViT-S features. Table 7 reports test Macro-$F1$ and Macro-AUROC over five random seeds. Performance is generally stable across a broad range of vocabulary sizes, but does not improve monotonically as K increases. Moreover, the optimal vocabulary size varies across datasets, encoders, and metrics: for example, $K = 24$ performs best on BRACS with ImageNet features, whereas $K = 48$ gives the highest Macro-$F1$ on CPTAC for both encoders. The results indicate that increasing the number of morphology tokens alone does not systematically im-

prove classification; rather, a moderate vocabulary is sufficient when the learned codewords capture class-relevant morphology.

Table 7: Effect of morphology vocabulary size on BRACS and CPTAC with two frozen feature encoders. We report test Macro-$F1$ and Macro-AUROC over five random seeds. All non-varied hyperparameters within each dataset–encoder setting are held fixed. **Bold** and underlined values denote the best and second-best results, respectively.

		BRACS		CPTAC	
Backbone	K	Macro-$F1$ ↑	Macro-AUROC ↑	Macro-$F1$ ↑	Macro-AUROC ↑
ResNet-18 ImageNet pretrained	4	$0.530_{\pm 0.052}$	$0.721_{\pm 0.046}$	$0.910_{\pm 0.004}$	$0.962_{\pm 0.003}$
	8	$0.528_{\pm 0.043}$	$0.744_{\pm 0.019}$	$\underline{0.911}_{\pm 0.004}$	$0.964_{\pm 0.006}$
	16	$0.537_{\pm 0.059}$	$0.746_{\pm 0.059}$	$0.907_{\pm 0.009}$	$\mathbf{0.965}_{\pm 0.004}$
	24	$\mathbf{0.573}_{\pm 0.034}$	$\underline{0.757}_{\pm 0.015}$	$0.911_{\pm 0.009}$	$0.964_{\pm 0.003}$
	32	$0.529_{\pm 0.043}$	$\mathbf{0.764}_{\pm 0.023}$	$\mathbf{0.914}_{\pm 0.011}$	$\mathbf{0.965}_{\pm 0.009}$
	48	$\underline{0.541}_{\pm 0.032}$	$0.734_{\pm 0.029}$	$\mathbf{0.914}_{\pm 0.007}$	$0.963_{\pm 0.003}$
	64	$0.529_{\pm 0.028}$	$0.728_{\pm 0.029}$	$0.907_{\pm 0.008}$	$0.963_{\pm 0.004}$
ViT-S SSL-pretrained	4	$0.681_{\pm 0.045}$	$0.860_{\pm 0.017}$	$0.933_{\pm 0.013}$	$0.981_{\pm 0.003}$
	8	$0.697_{\pm 0.046}$	$0.869_{\pm 0.018}$	$\underline{0.938}_{\pm 0.010}$	$\underline{0.983}_{\pm 0.001}$
	16	$\underline{0.717}_{\pm 0.012}$	$0.874_{\pm 0.011}$	$0.935_{\pm 0.013}$	$0.981_{\pm 0.002}$
	24	$\mathbf{0.722}_{\pm 0.030}$	$\underline{0.876}_{\pm 0.011}$	$0.936_{\pm 0.007}$	$0.982_{\pm 0.003}$
	32	$0.709_{\pm 0.020}$	$\mathbf{0.878}_{\pm 0.011}$	$0.935_{\pm 0.013}$	$0.981_{\pm 0.002}$
	48	$0.712_{\pm 0.017}$	$0.870_{\pm 0.018}$	$\mathbf{0.945}_{\pm 0.008}$	$\mathbf{0.983}_{\pm 0.001}$
	64	$0.701_{\pm 0.014}$	$0.867_{\pm 0.011}$	$0.937_{\pm 0.012}$	$0.981_{\pm 0.002}$

E.3 Soft-assignment temperature.

The assignment temperature γ_q controls the concentration of the soft morphology-token assignments in Equation 8. Table 8 reports two independent matched sweeps. On BRACS with ImageNet features, the intermediate temperature $\gamma_q = 0.35$ performs best, whereas a softer assignment at 0.50 is consistently worse. On CAMELYON16 with medical SSL features, $\gamma_q = 0.50$ gives the best joint result. These results support treating γ_q as a dataset/feature-dependent concentration parameter rather than expecting a single universally optimal temperature.

Table 8: Sensitivity to soft-assignment temperature γ_q under exactly matched configurations. **Bold** and underlined values denote the best and second-best results within each feature setting.

Dataset / Feature	γ_q	Macro-$F1$ ↑	Macro-AUROC ↑
BRACS / ImageNet	0.30	$\underline{0.517}_{\pm 0.072}$	$\underline{0.722}_{\pm 0.062}$
	0.35	$\mathbf{0.529}_{\pm 0.043}$	$\mathbf{0.764}_{\pm 0.023}$
	0.50	$0.521_{\pm 0.041}$	$0.720_{\pm 0.047}$
CAMELYON16 / Medical SSL	0.30	$0.947_{\pm 0.013}$	$\underline{0.980}_{\pm 0.007}$
	0.35	$\underline{0.950}_{\pm 0.016}$	$0.980_{\pm 0.008}$
	0.50	$\mathbf{0.951}_{\pm 0.011}$	$\mathbf{0.983}_{\pm 0.003}$

E.4 Tokenizer depth.

We examine the effect of tokenizer capacity by varying the depth of the trainable tokenizer projection f_θ from 1 to 3. The ablation is conducted on CAMELYON16 and CPTAC using both ImageNet-pretrained and medical SSL features, resulting in four dataset–feature settings. Within each setting, all remaining hyperparameters are held fixed so that only the tokenizer depth is varied.

As shown in Table 9, increasing tokenizer depth does not consistently improve classification performance. On CAMELYON16 with ImageNet features, depth 3 achieves the highest Macro-$F1$ and Macro-AUROC. With medical SSL features, depth 2 performs best on both metrics. On CPTAC with ImageNet features, depth 3 achieves the highest Macro-$F1$ and Macro-AUROC. In contrast, with medical SSL features, the shallow depth-1 tokenizer achieves the best performance on both metrics.

Overall, these results show that MELD does not require a deep tokenizer to construct an effective morphology-token representation. Performance is relatively insensitive to additional tokenizer

depth, and the preferred depth depends on the dataset and feature representation rather than increasing monotonically with model capacity.

Table 9: Effect of tokenizer depth across datasets and feature encoders. All remaining hyperparameters are fixed within each dataset–feature setting. Results are reported as mean ± standard deviation over five runs.

		Depth 1		Depth 2		Depth 3	
Dataset	Feature	Macro-$F1$	Macro-AUROC	Macro-$F1$	Macro-AUROC	Macro-$F1$	Macro-AUROC
CAMELYON16	ImageNet	$0.758_{\pm 0.043}$	$0.818_{\pm 0.033}$	$0.775_{\pm 0.034}$	$0.780_{\pm 0.035}$	$\mathbf{0.790}_{\pm 0.008}$	$\mathbf{0.820}_{\pm 0.023}$
CAMELYON16	Medical SSL	$0.943_{\pm 0.039}$	$0.970_{\pm 0.019}$	$\mathbf{0.951}_{\pm 0.011}$	$\mathbf{0.983}_{\pm 0.003}$	$0.947_{\pm 0.041}$	$0.974_{\pm 0.017}$
CPTAC	ImageNet	$0.897_{\pm 0.013}$	$0.952_{\pm 0.007}$	$0.911_{\pm 0.004}$	$0.958_{\pm 0.006}$	$\mathbf{0.914}_{\pm 0.011}$	$\mathbf{0.965}_{\pm 0.009}$
CPTAC	Medical SSL	$\mathbf{0.945}_{\pm 0.008}$	$\mathbf{0.983}_{\pm 0.001}$	$0.937_{\pm 0.010}$	$0.981_{\pm 0.002}$	$0.918_{\pm 0.015}$	$0.977_{\pm 0.003}$

The matched reruns show that MELD is not critically dependent on a large morphology vocabulary or a deep tokenizer, and that the assignment temperature admits a reasonably broad operating range once adapted to the feature representation. These sensitivity results complement the qualitative morphology and localization analyses by characterizing how the learned evidence representation behaves under controlled changes to its principal hyperparameters.

F Additional Codeword Assignment and PMI Visualizations

We provide additional visualizations of the learned morphology codewords and their class-specific PMI evidence on CAMELYON16 and BRACS. For this analysis, MELD is trained with $B = 6$ evidence-collection branches and $K = 32$ morphology codewords. After training, each tile is assigned to the codeword with the largest assignment probability. For each diagnostic class and codeword with assigned tiles,we randomly sample from that class eight tiles whose highest-probability assignment is to the codeword for pathologist review. A pathologist examines all eight tiles associated with each class–codeword pair and confirms whether they exhibit a consistent morphological pattern. To keep the visualization compact, we randomly select four of the eight reviewed tiles for display.

The retrieved tiles provide a qualitative view of the morphology partition induced by the learned tokenizer. Tiles assigned to the same codeword often exhibit similar tissue appearance and structural characteristics, whereas different codewords capture distinct recurring morphology patterns. Since the codewords are learned using only slide-level supervision, they should be interpreted as learned morphology tokens rather than predefined pathological categories.

For each diagnostic class, we additionally visualize its $B \times K$ signed PMI matrix below the codeword gallery, where rows correspond to evidence-collection branches and columns correspond to morphology codewords. Each entry $\boldsymbol{e}_{yb}[k]$ quantifies the class-specific diagnostic evidence associated with codeword k under branch b. Positive values indicate branch–codeword pairs that support the corresponding diagnosis, whereas negative values indicate evidence against it. The PMI matrices therefore reveal how the same learned morphology tokens can play different diagnostic roles across branches and classes.

Overall, the figures show both broadly shared morphology tokens and more class-discriminative tokens. Together, the codeword galleries and PMI matrices illustrate the two complementary aspects of MELD interpretation: the tokenizer characterizes *what* morphology is observed, while the signed PMI evidence indicates *how* that morphology supports or opposes each diagnosis.

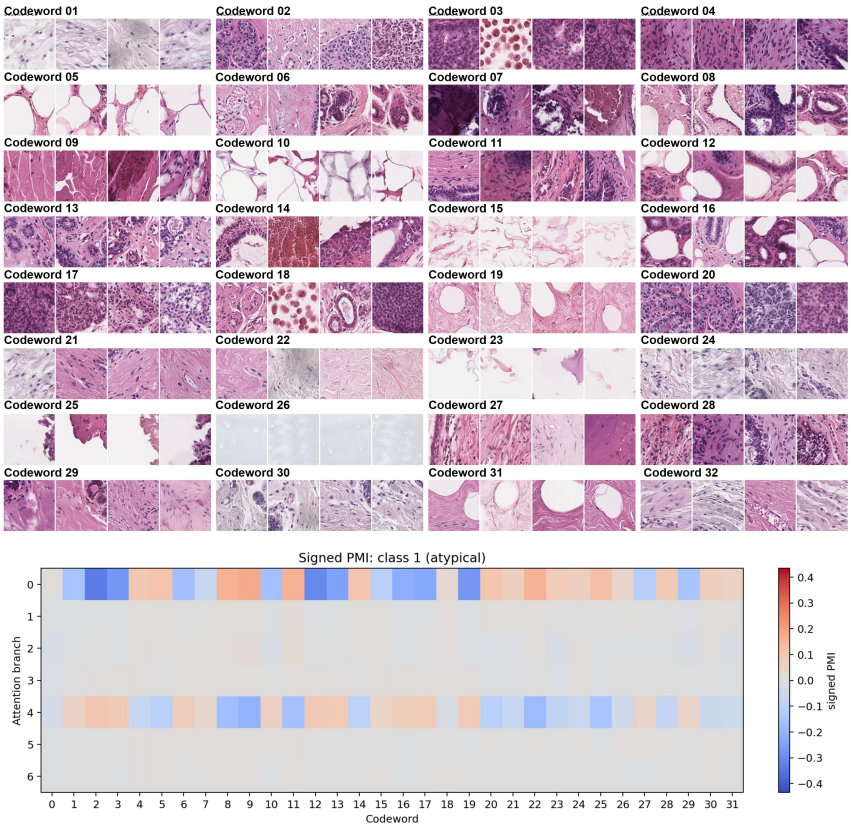

Figure 14: Codeword assignments and signed PMI evidence for the *atypical* class on BRACS.

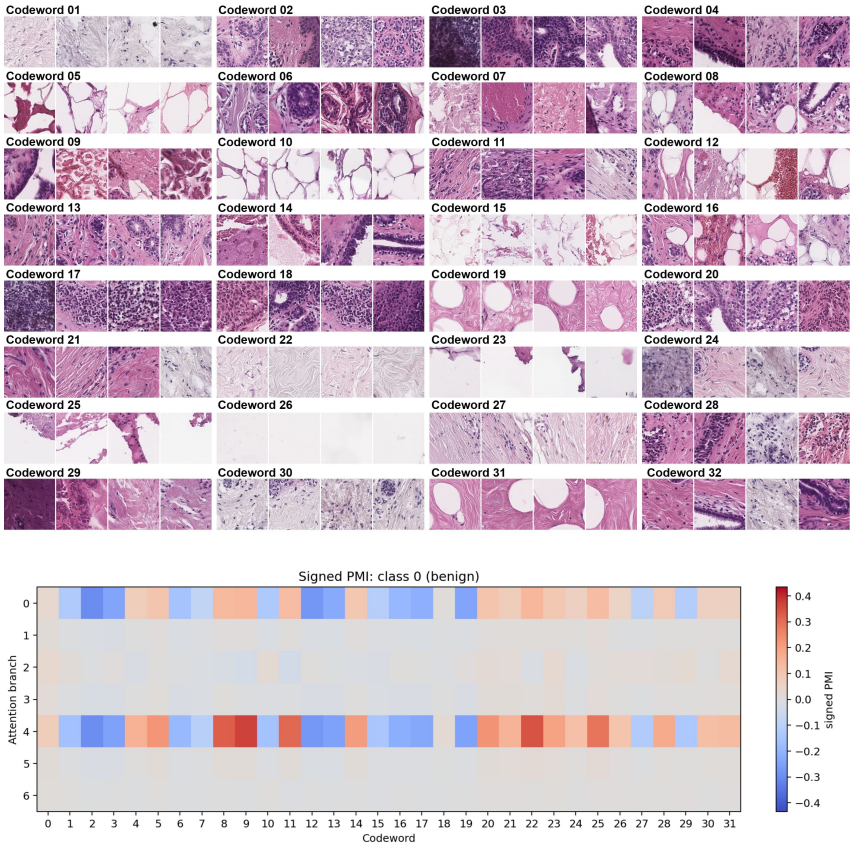

Figure 15: Codeword assignments and signed PMI evidence for the *benign* class on BRACS.

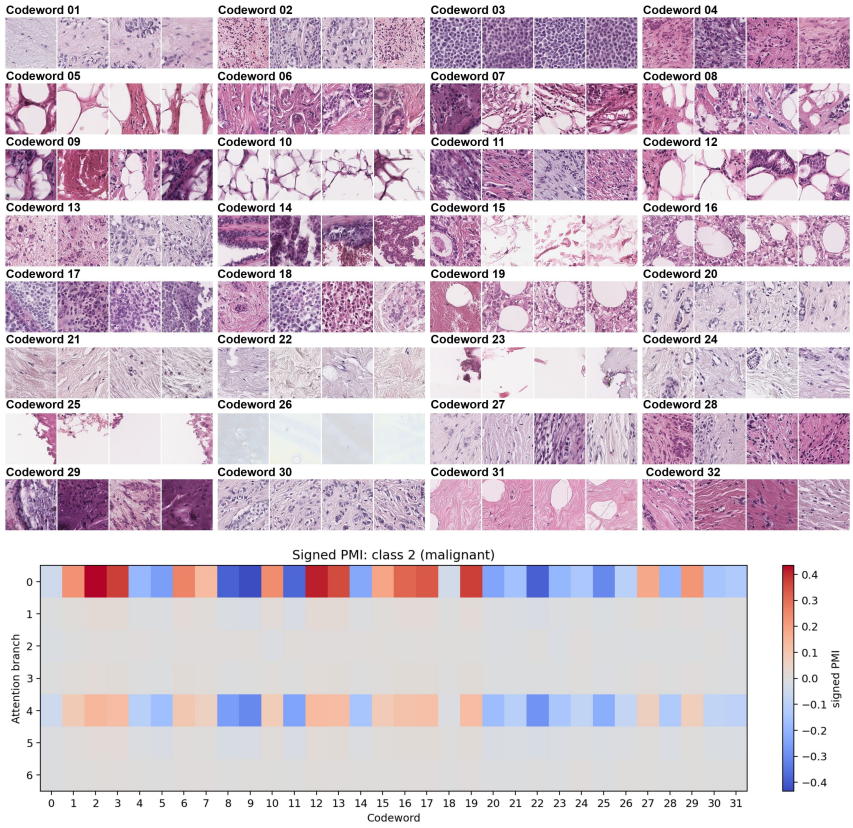

Figure 16: Codeword assignments and signed PMI evidence for the *malignant* class on BRACS.

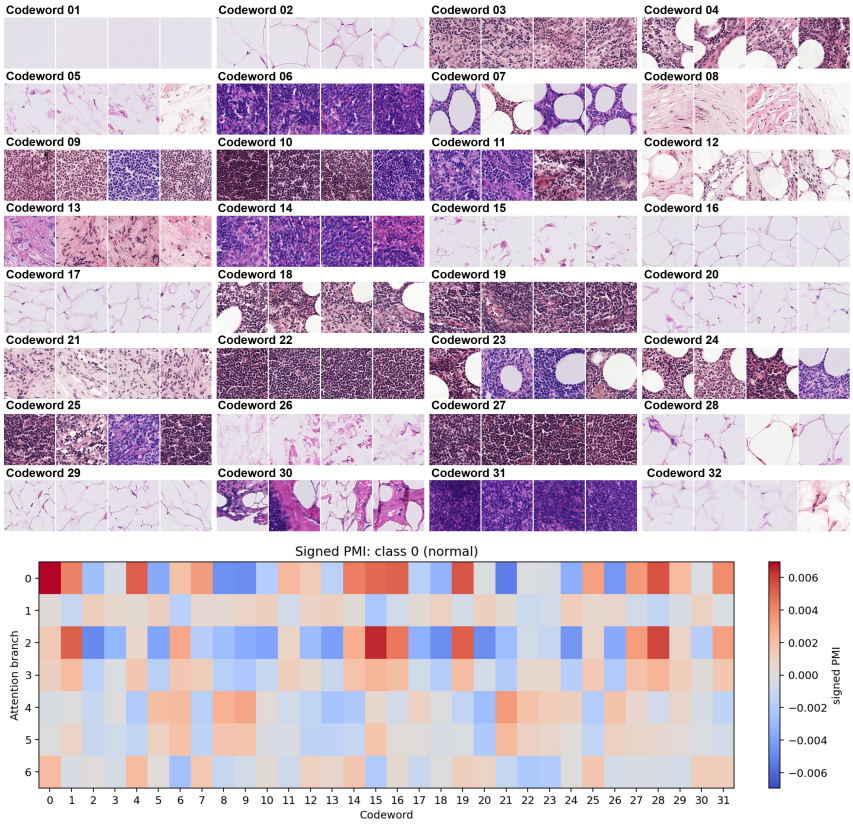

Figure 17: Codeword assignments and signed PMI evidence for the *normal* class on CAMELYON16.

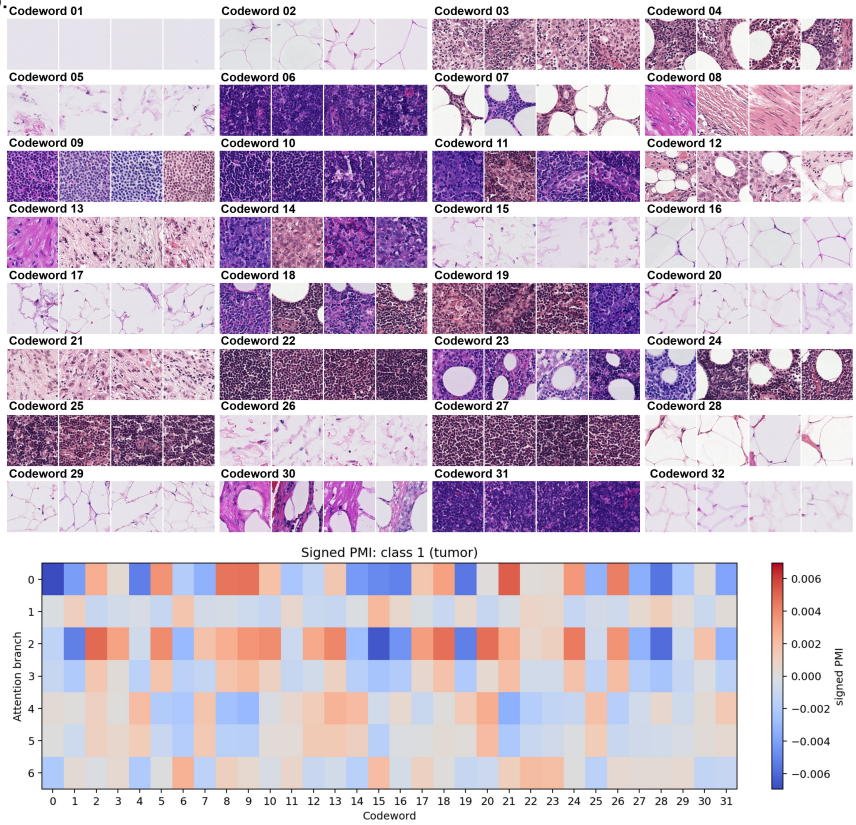

Figure 18: Codeword assignments and signed PMI evidence for the *tumor* class on CAMELYON16.

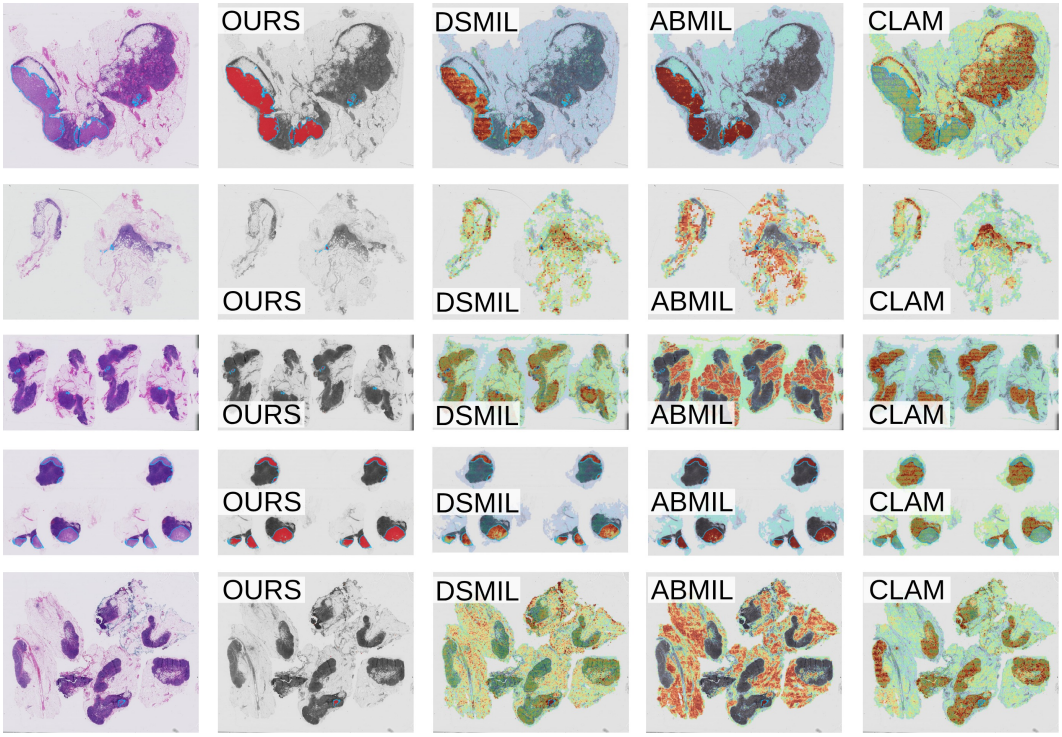

Figure 19: Representative tumor-localization examples on CAMELYON16 using medical SSL features. From left to right, we show the original WSI with the annotated tumor ROI, MELD's positive diagnostic-evidence map, and the attention maps of DSMIL, ABMIL, and CLAM-SB. Cyan contours indicate the pixel-level tumor annotations. For MELD, red regions denote positive PMI evidence for the tumor class, whereas the baseline heatmaps visualize their learned attention scores. MELD's positive evidence is more tightly concentrated within the annotated tumor regions while suppressing responses in non-tumor tissue.

G Localization Using Evidence Maps

We further investigate whether MELD's class-specific diagnostic evidence can localize tumor regions. Unlike attention, which indicates where information is collected, PMI-based evidence additionally specifies whether the morphology observed at each location supports or opposes a target diagnosis. We construct a class-specific *evidence map* by projecting the PMI values of the target class back onto individual tiles according to their morphology-token assignments. Tiles with positive projected PMI are regarded as positive diagnostic-evidence regions for the target class.

G.1 Quantitative Localization on CAMELYON16

CAMELYON16 provides pixel-level annotations of metastatic regions, enabling direct quantitative evaluation of tumor localization. For each tumor-positive CAMELYON16 test slide, we use the tumor class as the target. All baseline models are trained using their standard implementations and default model-specific settings, without modifying their architectures or training objectives for localization. To provide a controlled comparison of localization signals commonly available in attention-based MIL, we use the attention scores produced by each trained baseline as its localization signal. The per-tile attention scores are rescaled to $[0, 1]$ within each slide and thresholded at 0.5 to obtain binary localization masks. For MELD, we instead use its native class-specific diagnostic evidence, which is directly defined by the classification model. For MELD, we project the class-specific PMI evidence for the tumor class back to individual tiles and classify a tile as tumor-supporting when $r_{si}^{\text{tumor}} > 0$. The resulting binary masks are compared with the CAMELYON16 tumor annotations.

Table 10: Tumor localization on CAMELYON16 using the standard implementations of the compared MIL methods. MELD uses its native class-specific positive evidence, whereas the baseline models use their native attention scores as the common attention-based localization signal. **Bold** and underlined values denote the best and second-best results, respectively.

Method	Dice ↑	Specificity ↑
CLAM-SB Nature 2021	0.0660	0.5035
ABMIL ICML 2018	0.2329	0.7117
DSMIL CVPR 2021	0.3009	0.6919
IBMIL CVPR 2023	0.2675	0.7065
ASMIL ICLR 2026	<u>0.3942</u>	<u>0.8439</u>
MELD	**0.5288**	**0.9936**

Table 11: Localization performance of isolated tumor cells (ITCs) on CAMELYON17 using medical SSL features. MELD uses its positive ITC-class diagnostic-evidence map, whereas the MIL baselines use their attention maps as localization signals. **Bold** and underlined values denote the best and second-best results, respectively.

Method	Dice ↑	Specificity ↑
CLAM-MB Nature 2021	0.0340	0.5338
ABMIL ICML 2018	0.0347	<u>0.5888</u>
DSMIL CVPR 2021	0.0148	0.4500
IBMIL CVPR 2023	<u>0.0623</u>	0.5601
ASMIL ICLR 2026	0.0563	0.5647
MELD	**0.1688**	**0.9997**

We report Dice to measure overlap with the annotated tumor region and specificity to measure rejection of non-tumor tissue.

As shown in Table 10, MELD achieves the best performance on both metrics, with a Dice score of 0.5288 and a specificity of 0.9936. The strongest baseline, ASMIL, obtains a Dice score of 0.3942 and a specificity of 0.8439 . Thus, MELD improves tumor-region overlap while simultaneously reducing localization on non-tumor tissue. These results indicate that class-specific positive evidence provides a localization signal that is more closely aligned with pathological tumor annotations than attention alone.

Figure 19 provides representative qualitative comparisons. Baseline attention maps often respond broadly across diagnostically informative tissue, including regions outside the annotated tumor ROI. In contrast, MELD concentrates positive tumor evidence within the metastatic regions while assigning little or no positive evidence to surrounding tissue. This difference reflects the distinct semantics of the two signals: attention measures how strongly a region participates in evidence collection, whereas signed PMI determines whether its observed morphology supports or opposes the tumor diagnosis.

G.2 Localization of Isolated Tumor Cells on CAMELYON17

In CAMELYON17, isolated tumor cells (ITCs) are defined as single tumor cells or tumor-cell clusters no larger than $0.2\,\mathrm{mm}$ or containing fewer than 200 cells. Because ITCs occupy only a minute fraction of a whole-slide image, their diagnostic signal can be easily overwhelmed by benign tissue. Their localization therefore provides a particularly challenging test of whether a MIL model can preserve highly sparse tumor evidence.

As shown in Figure 20, MELD successfully localizes the minute ITC foci across these challenging CAMELYON17 cases. Although the metastatic regions occupy only a tiny fraction of each WSI, the corresponding tiles receive strong positive ITC evidence. In contrast, the attention maps of DSMIL, ABMIL, and CLAM are substantially more diffuse or emphasize regions outside the ITC focus, failing to reliably isolate these sparse tumor deposits. These examples indicate that MELD can preserve diagnostically relevant evidence even when the target morphology is extremely sparse.

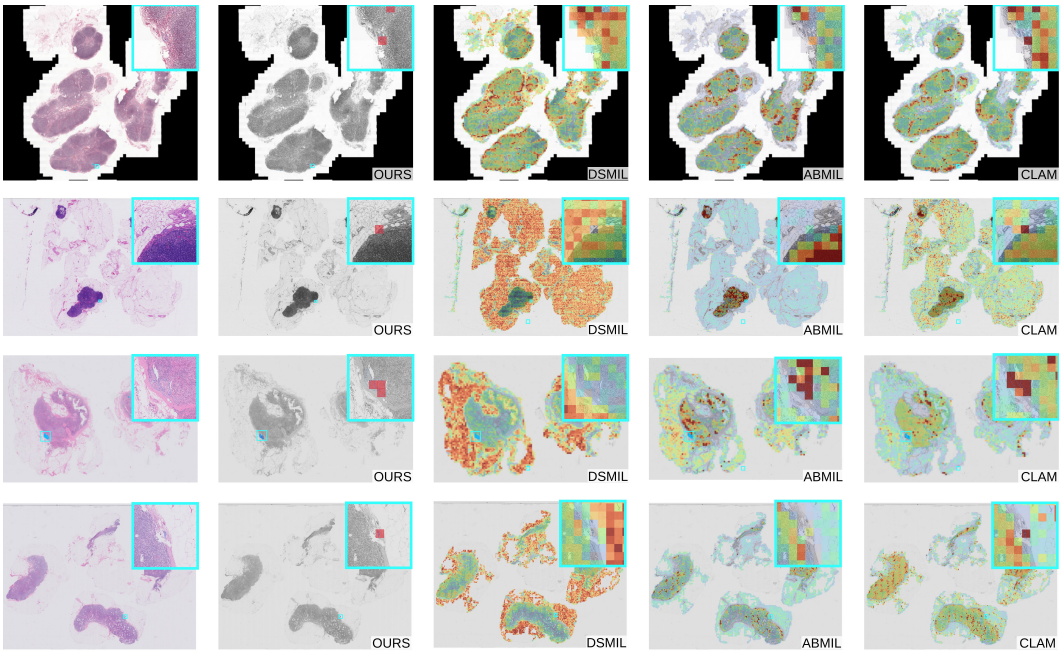

Figure 20: **Localization of isolated tumor cells (ITCs) on CAMELYON17.** Each row shows an ITC-positive WSI, with the cyan inset magnifying the corresponding ITC region. From left to right, we show the H&E slide, MELD's positive signed-PMI evidence for the ITC class, and the attention maps produced by DSMIL, ABMIL, and CLAM. Despite the extremely small extent of the metastatic foci, MELD consistently assigns positive ITC-class evidence to the ITC regions. In contrast, the attention-based MIL methods being compared produce diffuse or spatially misaligned responses and fail to reliably identify these minute lesions. Among the compared methods, MELD is the only one that successfully localizes the ITC foci across these examples.

Together with the quantitative localization results on CAMELYON16, these observations highlight the distinction between MELD's signed evidence and conventional MIL attention. Attention indicates *where* a model collects information, but does not specify whether the observed morphology supports or opposes a particular diagnosis. MELD instead assigns each local morphology signed, class-specific diagnostic evidence. Positive ITC-class evidence therefore provides a direct localization signal for ITC-associated morphology, enabling MELD to identify even highly sparse lesions such as ITCs.

H LIMITATIONS

MELD characterizes which morphology tokens are present in a WSI and how they contribute to a diagnostic decision, but it does not explicitly model how these tokens are spatially organized across the slide. The slide-level morphology representation is obtained by aggregating tile-level morphology assignments under the attention distribution and marginalizing over tile identity. Consequently, inter-tile adjacency and relative spatial arrangement are not explicitly preserved. Slides with similar attention-weighted morphology distributions but different tissue organizations may therefore receive similar representations. This limitation concerns spatial relationships between tiles. Local morphological structure within individual tiles can still be represented by the feature encoder, and MELD's tile-level evidence can be mapped back to WSI coordinates for visualization. Incorporating spatial interactions among morphology tokens while retaining MELD's explicit probabilistic evidence structure is a promising direction for future work.