ONE TILE, MULTIPLE INSTANCES: RETHINKING MIL FOR SPARSE DIAGNOSTIC EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

In weakly supervised Whole Slide Image (WSI) classification, feature extractors typically compress each image tile into a single global embedding. Consequently, slide-level aggregators are restricted to this coarse tile scale, concealing fine-grained sub-tile evidence from the attention mechanism. We introduce DI-MIL, a framework that decouples encoding context from instance granularity through decomposed instances. By clustering dense spatial tokens from a frozen foundation model within each tile, DI-MIL converts a single tile into multiple independently weighted instance embeddings. As a training-free post-encoding module, DI-MIL integrates seamlessly into existing pipelines without requiring re-encoding or downstream architectural modifications. We evaluate DI-MIL on cytopathology, a challenging testbed where sparse diagnostic signals are easily diluted within standard tiles. Across four datasets, three frozen foundation models, and two attention-based aggregators, DI-MIL demonstrates consistent efficacy, improving 67 of 72 metric-level comparisons, with the largest mean gains reaching 3.64 points under cytopathology-specific backbones. In a broader comparison against seven representative MIL baselines, DI-MIL paired with ACMIL achieves highest mean performance in 33 of 36 backbone-dataset-metric comparisons. Ablations show that direct smaller tiling inflates the extracted tile count by up to 43.3$\times$ with non-monotonic performance, whereas DI-MIL incurs zero additional image-extraction overhead while achieving the strongest overall results. These results establish instance construction as an orthogonal design dimension in MIL, supporting DI-MIL as a cost-efficient solution under sparse diagnostic evidence. Code will be released upon acceptance.

1 INTRODUCTION

Deep learning-based pathology diagnosis on whole-slide images (WSIs) presents a classic weak-supervision problem, where sparse visual evidence determines the slide-level label. Multiple instance learning (MIL) addresses this by treating a slide as a bag of instances aggregated, typically via attention mechanisms under slide-level supervision (Ilse et al., 2018; Li et al., 2021; Shao et al., 2021). Typically, standard MIL workflows adopt a two-stage paradigm: image tiles are extracted from WSIs and encoded into fixed feature embeddings by a pre-trained backbone, which are subsequently aggregated into a slide-level representation. This standard pipeline compresses each tile into a single global embedding, rigidly coupling aggregation granularity to the feature extractor's receptive field and forcing the model to treat the tile as indivisible.

However, this coupling poses a limitation when diagnostic evidence is sparse within a tile. Cytopathology provides a concrete example: scattered malignant cells are often interspersed among abundant normal cells and non-diagnostic background (Dov et al., 2021; Butke et al., 2021; Zhao et al., 2024). In such cases, a tile can be diagnostically decisive even if only a minute fraction of its region is relevant. Because each tile is represented as a single embedding, attention mechanisms can only assign a single weight to the tile as a whole. As a result, slide attention cannot weight critical sub-tile morphology independently of surrounding background, conflating tile identification with sub-tile evidence weighting.

To address this, we argue that the spatial encoding context of the feature extractor should be decoupled from the instance granularity used in slide-level aggregation. A straightforward workaround

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.

is to extract smaller source tiles; however, this presents an inherent trade-off: while smaller tiles increase aggregation granularity, they severely shrink the spatial context provided to the backbone. Vision foundation models offer a way out: their dense spatial tokens preserve fine-grained localized semantics while maintaining a broad contextual field of view (Siméoni et al., 2025; Li et al., 2025). This capability challenges the traditional MIL assumption, raising our core question: *must features encoded together within a tile also be aggregated as a single instance?* We posit that the source tile should remain the unit of contextual encoding, while its sub-tile spatial tokens serve as individual instances for slide-level aggregation.

To this end, we introduce **DI-MIL**, a feature-level framework that achieves this separation via decomposed instances. For each tile, DI-MIL clusters ℓ_2-normalized spatial tokens by feature similarity, averages the original token representations within each cluster, and applies layer normalization to the pooled cluster embeddings. Each cluster embedding enters the slide bag as an independent instance, allowing sub-tile components to receive distinct attention weights. By expanding a slide bag of N source tiles into kN fine-grained instances, DI-MIL enriches instance granularity without requiring additional image crops or extra encoder forward passes.

Using cytopathology as a testbed for sparse evidence, we systematically evaluate DI-MIL as a plug-and-play enhancement across a full matrix of 4 datasets, 3 vision foundation backbones, and 2 MIL aggregators. Across 72 total evaluation settings combining multiple datasets, backbones, aggregators, and metrics, DI-MIL improves performance in 67 configurations, where maximum mean gains reaching 3.64 points under cytopathology-specific backbones. In a broader comparison against seven representative MIL baselines, DI-MIL combined with ACMIL achieves the highest mean result in 33 of 36 backbone-dataset-metric comparisons. Ablation studies establish $k = 4$ as a default setting and demonstrate that extracting smaller source tiles substantially increases computational and storage overhead without matching the performance achieved by DI-MIL.

In summary, our main contributions are threefold:

- We delineate the distinction between encoding context and instance granularity in MIL, uncovering the structural limitation of single-embedding tiles when diagnostic evidence is sparse within a tile.
- We propose DI-MIL, a framework that decomposes dense spatial tokens into independently weighted instances, refining aggregation granularity without requiring extra encoder forward passes or cell-level annotations.
- We extensively evaluate DI-MIL on cytopathology benchmarks, demonstrating that decoupling encoding context from instance granularity outperforms both standard MIL baselines and naive extraction of smaller source tiles.

2 Related Work

Aggregation in multi-instance learning. Multiple instance learning (MIL) algorithms for whole-slide image (WSI) classification have primarily focused on downstream aggregation mechanisms, including attention pooling (Ilse et al., 2018), instance-relation modeling (Li et al., 2021; Shao et al., 2021), pseudo-bag distillation (Zhang et al., 2022), and long-sequence modeling (Yang et al., 2024). Subsequent advances such as ACMIL and ASMIL refine attention calibration during pooling (Zhang et al., 2024; Ye et al., 2026), while HMIL introduces hierarchical alignment across scales (Jin et al., 2025). While existing aggregators innovate in feature grouping and weighting, they treat each tile as a monolithic vector. DI-MIL decomposes dense spatial tokens into fine-grained instances prior to slide-level aggregation, making it fully complementary to downstream pooling architectures.

Dense representations in WSI learning. Providing such fine-grained inputs relies on dense feature representations. Modern vision foundation models extract dense token features alongside global image representations, providing fine-grained spatial information (Siméoni et al., 2025). In computational pathology, methods such as PathVQ retain dense spatial tokens via vector quantization (Li et al., 2025). However, PathVQ compresses intra-tile spatial tokens into a single tile representation prior to slide-level pooling. In contrast, DI-MIL preserves multiple cluster embeddings per tile as separate inputs throughout the MIL aggregation stage to prevent localized pathology from being diluted prior to slide-level weighting.

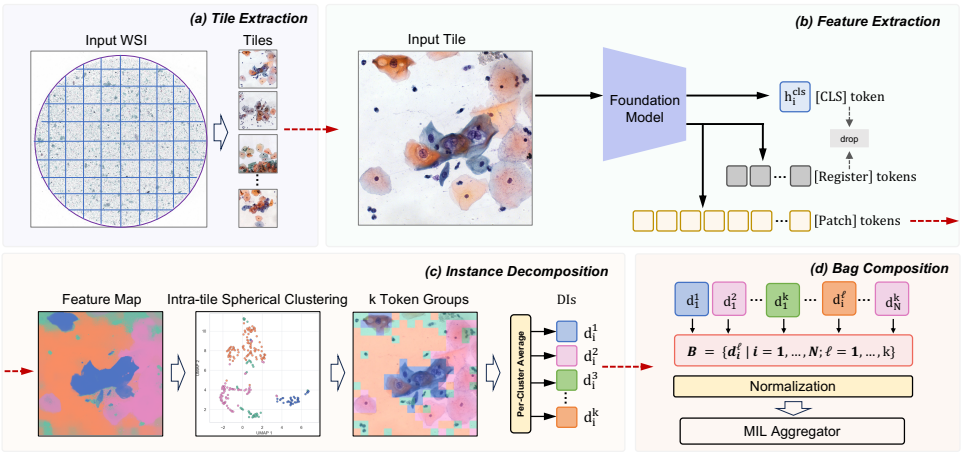


Figure 1: **Overview of DI-MIL. (a)** A whole-slide cytopathology image is partitioned into source tiles. **(b)** A frozen transformer-based foundation model extracts a global [CLS] representation and dense patch-token features from each tile. **(c)** DI-MIL performs intra-tile spherical clustering on dense token representations, aggregates tokens within each cluster, and constructs multiple decomposed instances per tile. **(d)** The resulting decomposed instances from all tiles are assembled into the slide-level MIL bag and processed by a downstream MIL aggregator.

Token clustering and prototype construction. To structure and condense such dense spatial tokens, token clustering is widely employed. Organizing spatial representations through clustering has been explored across various stages of model pipelines. For instance, TCFormer progressively merges tokens inside the visual encoder to adapt spatial receptive fields (Zeng et al., 2022). Conversely, PTCMIL applies prompt-guided clustering over global tile representations across the whole slide to construct slide-level prototypes (Zhao et al., 2026). Distinct from these approaches, DI-MIL performs intra-tile spherical clustering independently on dense tokens after frozen feature extraction. This choice dynamically reduces token redundancy while strictly preserving local spatial context, avoiding the computational complexity of slide-wide clustering.

Cytopathology and sparse diagnostic evidence. The need for sub-tile instance decomposition is most critical in cytopathology, where evidence sparsity is extreme. Weakly supervised learning in cytopathology addresses severe evidence sparsity across various clinical applications, including thyroid malignancy prediction (Dov et al., 2021), urothelial carcinoma classification (Butke et al., 2021), and malignant effusion classification (Kim et al., 2025). Domain-specific architectures such as LESS address irregular cell distributions via multi-scale modeling (Zhao et al., 2024), while frameworks such as Smart-CCS and UniCAS leverage specialized foundation model pretraining (Jiang et al., 2025; 2026). These works establish cytopathology as a benchmark for learning under sparse diagnostic evidence. We utilize this setting to evaluate how intra-tile dense tokens should be structured into MIL instances when informative content represents only a fraction of the tile area.

3 Decomposed-Instance Multiple Instance Learning

Conventional WSI multiple-instance learning (MIL) uses the encoder tile as the basic unit of slide-level aggregation. This couples two distinct design choices: the spatial context seen by the image encoder and the granularity at which the MIL model can weight diagnostic evidence. Using smaller tiles can produce a finer-grained bag, but simultaneously changes the encoder input and increases the number of image crops and encoder evaluations.

We introduce **Decomposed-Instance MIL (DI-MIL)** to decouple these two factors. DI-MIL preserves the original tile extraction and encoder field of view, while decomposing the patch-token representations already produced for each tile into multiple MIL instances. The slide-level model can therefore operate at a finer instance granularity without re-encoding the WSI at a smaller tile size. An overview is shown in Figure 1.

3.1 PRELIMINARIES

Let a whole-slide image X be partitioned into N non-overlapping image tiles $\{p_i\}_{i=1}^{N}$. A frozen transformer encoder f processes each tile independently. We denote the final-layer class-token representation by $\boldsymbol{h}_i^{\mathrm{cls}} \in \mathbb{R}^D$ and the corresponding M patch-token representations by $\{\boldsymbol{h}_{ij}\}_{j=1}^{M}$, where $\boldsymbol{h}_{ij} \in \mathbb{R}^D$. Thus,

$$f(p_i) = \left(\boldsymbol{h}_i^{\mathrm{cls}}, \{\boldsymbol{h}_{ij}\}_{j=1}^{M}\right). \tag{1}$$

Conventional WSI MIL retains only the class-token representation of each tile, forming the slide-level bag

$$\mathcal{B}_{\mathrm{tile}} = \{\boldsymbol{h}_i^{\mathrm{cls}}\}_{i=1}^{N}, \qquad \hat{y} = g_\theta(\mathcal{B}_{\mathrm{tile}}), \tag{2}$$

where g_θ is an attention-based MIL aggregator.

Equating the MIL instance to the tile-level embedding $\boldsymbol{h}_i^{\mathrm{cls}}$ restricts instance granularity to the physical tile. Consequently, visual patterns that are distinct within the same tile cannot receive separate MIL attention weights, although they remain represented by different patch-token features within the encoder output.

We refer to the amount of encoded visual content summarized by one independently weighted MIL instance as *instance granularity*. Conventional MIL fixes this granularity at the tile level. DI-MIL instead constructs multiple MIL instances from the patch-token representations of each tile while keeping the encoder input unchanged.

3.2 DECOMPOSED-INSTANCE CONSTRUCTION

In cytopathology, an extracted tile p_i exhibits severe intra-tile heterogeneity, often blending sparse diagnostically abnormal cells with abundant benign background. Compressing p_i into a single global embedding $\boldsymbol{h}_i^{\mathrm{cls}}$ forces the aggregator to assign a single attention weight to this mixed representation, diluting localized diagnostic evidence. By contrast, for each tile p_i, DI-MIL replaces its single tile-level representation $\boldsymbol{h}_i^{\mathrm{cls}}$ with k decomposed instances constructed from the corresponding patch-token representations $\{\boldsymbol{h}_{ij}\}_{j=1}^{M}$. As illustrated in Fig. 1, while the frozen encoder outputs both [CLS] token and patch tokens, DI-MIL bypasses the global class-token and restructures only the patch-token representations into slide-level MIL instances without altering the upstream forward pass.

Intra-tile token clustering. For each tile p_i, we group its M patch-token representations by cosine similarity in feature space. We first ℓ_2-normalize each representation as $\bar{\boldsymbol{h}}_{ij} = \boldsymbol{h}_{ij}/\|\boldsymbol{h}_{ij}\|_2$ and apply spherical k-means independently per tile. Denoting $\boldsymbol{\mu}_{i\ell}$ as the unit-norm centroid of the ℓ-th cluster, each patch representation is assigned to its nearest centroid:

$$a_{ij} = \arg\max_{\ell \in \{1,\dots,k\}} \bar{\boldsymbol{h}}_{ij}^{\top} \boldsymbol{\mu}_{i\ell}, \qquad \mathcal{C}_{i\ell} = \{j \mid a_{ij} = \ell\}, \qquad \|\boldsymbol{\mu}_{i\ell}\|_2 = 1. \tag{3}$$

This yields k disjoint token groups $\{\mathcal{C}_{i1}, \dots, \mathcal{C}_{ik}\}$ for each tile. Because clustering is performed independently in feature space, it requires neither spatial contiguity among patches within a cluster nor cross-tile cluster alignment.

Decomposed-instance pooling. Each cluster is converted into a decomposed instance by averaging the original patch representations assigned to it, followed by layer normalization:

$$\boldsymbol{d}_i^{\ell} = \mathrm{LN}\left(\frac{1}{|\mathcal{C}_{i\ell}|} \sum_{j \in \mathcal{C}_{i\ell}} \boldsymbol{h}_{ij}\right), \qquad \ell = 1, \dots, k. \tag{4}$$

Note that the normalized representations $\bar{\boldsymbol{h}}_{ij}$ are used strictly for cluster assignment, whereas the final instances $\boldsymbol{d}_i^{\ell}$ are pooled from the original encoder representations $\boldsymbol{h}_{ij}$. Consequently, each tile contributes k decomposed instances $\{\boldsymbol{d}_i^{\ell}\}_{\ell=1}^{k}$ instead of a single global vector, yielding the slide-level DI bag across all N tiles:

$$\mathcal{B}_{\mathrm{DI}} = \left\{\boldsymbol{d}_i^{\ell} \mid i = 1, \dots, N,\ \ell = 1, \dots, k\right\}, \qquad |\mathcal{B}_{\mathrm{DI}}| = kN. \tag{5}$$

In effect, DI-MIL enriches the instance granularity exposed to slide-level aggregation while leaving both the source image tiles and the upstream visual encoding computation entirely untouched.

3.3 MIL Integration and Computational Overhead

By aforementioned construction, each decomposed instance $\boldsymbol{d}_i^\ell$ shares the same feature dimensionality as the original tile-level representation $\boldsymbol{h}_i^{\text{cls}}$. This allows the expanded bag $\mathcal{B}_{\text{DI}}$ to be seamlessly processed by standard attention-based MIL aggregators without any architectural modifications:

$$\hat{y} = g_\theta(\mathcal{B}_{\text{DI}}). \tag{6}$$

For permutation-invariant aggregators, the local cluster index ℓ requires no cross-tile alignment; each $\boldsymbol{d}_i^\ell$ is simply treated as an independent instance within the slide-level bag.

In summary, DI-MIL alters instance granularity purely at the feature level, expanding the slide-level bag size from N to kN without modifying the image-encoding stage or requiring additional forward passes. Given pre-extracted tokens, decomposed instances are constructed once and cached prior to training, incurring minimal computational overhead compared to visual re-encoding.

4 Experiments

4.1 Experimental Setup

We evaluate four binary cytopathology WSI datasets: *Cervical* (978 slides), *Serous Fluid* (2,717 slides), *Respiratory* (1,041 slides), and the public *Oral2026* dataset (363 slides) (Jain et al., 2026) released on Zendono.[1]. For the three institutional cohorts, non-overlapping 1200×1200-pixel tiles are extracted after Otsu-based background filtering. *Oral2026* uses 2048×2048-pixel level-0 tiles following the established protocol. The institutional cohorts use fixed 10-fold cross-validation, whereas *Oral2026* uses 5-fold cross-validation following previous work (Mukherjee et al., 2025). We report metrics across folds including AUC, accuracy (ACC), and macro-F1.

Backbones, MIL aggregators, and training. Feature extraction uses one general-domain foundation model, DINOv3 (Siméoni et al., 2025), alongside two cytopathology-specific foundation models, Smart-CCS (Jiang et al., 2025) and UniCAS (Jiang et al., 2026). Evaluations apply DI-MIL directly to two set-based attention aggregators: ABMIL (Ilse et al., 2018) and ACMIL (Zhang et al., 2024). Sequence-based and spatial-structure MIL architectures serve as external reference baselines. All experiments use frozen 1024-dimensional embeddings, a learning rate of 2×10^{-4}, a batch size of one WSI, and early stopping within 100 epochs. DI-MIL uses $k = 4$ by default. Within each paired comparison, the extracted features, data splits, and training protocol are held fixed. Full settings are provided in Appendix A.1.

4.2 Main Result: Performance across Backbones and Aggregators

Table 1 summarizes the performance of DI-MIL across three frozen backbones and two aggregators while holding downstream architectures constant within each paired comparison. Replacing conventional single-embedding bags with DI-MIL yields performance improvements in 67 out of 72 evaluation settings. Aggregate improvements are consistently positive across datasets, ranging from +1.10 to +3.21 percentage points. Fold-paired analyses in Appendix A.6 further characterize the variability of these gains across matched folds. The mean performance gains are most pronounced in liquid-based and *Respiratory* cohorts, *Serous Fluid* and *Respiratory*, where rare diagnostic cells are diluted by abundant background. Decoupling instance granularity enables slide attention to focus on evidence-enriched sub-tile clusters. Across feature extractors, mean performance gains are highest under Smart-CCS (+3.64 points), followed by UniCAS (+1.59 points) and DINOv3 (+0.51 points). This trend indicates that DI-MIL leverages fine-grained representations more effectively when dense tokens capture domain-specific cytomorphological details. We further analyze this backbone-dependent behavior through quantitative and qualitative dense-token analyses in Appendix A.2. Furthermore, performance gains remain consistent between ABMIL (+1.86 points) and ACMIL (+1.88 points), demonstrating transferability across distinct attention aggregators.

4.3 Comparison with Representative MIL Baselines

Decoupling instance construction proves competitive with complex architectural modifications to downstream aggregators. Table 2 compares DI-MIL paired with ACMIL against seven representative MIL baselines under identical backbone features and data splits. DI-MIL with ACMIL achieves

[1] https://doi.org/10.5281/zenodo.20643817

Table 1: Comparison between Tile instances and our Decomposed instances (DI) across 4 cytopathology datasets, 3 foundation-model backbones, and 2 MIL aggregators. All methods use identical downstream MIL architectures and training protocols. Results are reported as mean with standard deviation in parentheses. Δ denotes the absolute performance change from Tile to DI.

Dataset	Instance	ABMIL			ACMIL		
		AUC	ACC	F1	AUC	ACC	F1
DINOv3							
Cervical	Tile	91.40 (2.11)	84.56 (3.45)	84.40 (3.44)	93.83 (1.96)	87.22 (3.33)	87.01 (3.43)
	DI (Ours)	91.52 (2.65)	85.69 (3.46)	85.48 (3.59)	93.86 (1.97)	88.24 (3.12)	88.06 (3.23)
	Δ	+0.12	+1.13	+1.08	+0.03	+1.02	+1.05
Serous Fluid	Tile	85.37 (2.57)	78.76 (2.23)	78.70 (2.27)	86.11 (3.49)	79.13 (3.21)	79.11 (3.21)
	DI (Ours)	85.80 (3.58)	79.53 (3.45)	79.49 (3.49)	86.79 (3.50)	80.09 (3.56)	80.06 (3.56)
	Δ	+0.43	+0.77	+0.79	+0.68	+0.96	+0.95
Respiratory	Tile	79.38 (4.18)	74.73 (4.14)	70.64 (5.31)	81.05 (4.66)	75.98 (3.15)	72.11 (3.87)
	DI (Ours)	80.77 (5.35)	76.37 (4.50)	71.58 (7.91)	82.01 (4.92)	76.75 (4.19)	72.67 (5.70)
	Δ	+1.39	+1.64	+0.94	+0.96	+0.77	+0.56
Oral2026	Tile	88.06 (5.07)	79.84 (2.85)	70.93 (5.04)	88.37 (5.15)	80.77 (2.89)	71.45 (5.90)
	DI (Ours)	85.72 (6.30)	82.61 (3.01)	66.26 (12.28)	89.36 (4.53)	79.91 (4.09)	72.47 (5.20)
	Δ	-2.34	+2.77	-4.67	+0.99	-0.86	+1.02
Smart-CCS							
Cervical	Tile	94.38 (1.58)	87.42 (2.41)	87.28 (2.53)	94.31 (1.90)	86.82 (3.36)	86.61 (3.48)
	DI (Ours)	94.71 (1.95)	90.08 (1.93)	89.99 (1.95)	95.89 (1.21)	89.98 (3.36)	89.87 (3.35)
	Δ	+0.33	+2.66	+2.71	+1.58	+3.16	+3.26
Serous Fluid	Tile	84.22 (2.22)	76.52 (2.74)	76.49 (2.74)	85.04 (2.82)	77.51 (3.20)	77.43 (3.25)
	DI (Ours)	87.63 (2.85)	80.05 (2.54)	79.98 (2.61)	88.33 (2.09)	80.64 (2.62)	80.60 (2.63)
	Δ	+3.41	+3.53	+3.49	+3.29	+3.13	+3.17
Respiratory	Tile	79.36 (4.13)	74.16 (3.11)	70.04 (4.38)	81.21 (3.45)	76.17 (3.19)	71.83 (4.47)
	DI (Ours)	87.50 (2.30)	80.11 (2.14)	77.25 (2.79)	88.85 (2.07)	82.32 (3.10)	80.08 (3.48)
	Δ	+8.14	+5.95	+7.21	+7.64	+6.15	+8.25
Oral2026	Tile	83.91 (5.38)	83.18 (2.74)	67.18 (13.01)	88.20 (5.44)	81.73 (6.14)	72.26 (4.25)
	DI (Ours)	84.17 (6.64)	81.88 (3.13)	71.26 (6.45)	89.13 (4.86)	84.52 (5.25)	75.69 (8.27)
	Δ	+0.26	-1.30	+4.08	+0.93	+2.79	+3.43
UniCAS							
Cervical	Tile	93.68 (2.43)	88.66 (2.36)	88.54 (2.39)	95.42 (0.74)	89.06 (1.73)	88.84 (1.85)
	DI (Ours)	94.35 (1.19)	89.27 (2.35)	89.15 (2.40)	95.79 (1.62)	89.57 (2.14)	89.39 (2.23)
	Δ	+0.67	+0.61	+0.61	+0.37	+0.51	+0.55
Serous Fluid	Tile	84.04 (2.21)	76.85 (2.04)	76.81 (2.01)	84.45 (2.40)	76.52 (2.27)	76.48 (2.26)
	DI (Ours)	85.13 (3.09)	79.10 (2.69)	79.07 (2.70)	86.06 (2.41)	78.43 (2.04)	78.39 (2.01)
	Δ	+1.09	+2.25	+2.26	+1.61	+1.91	+1.91
Respiratory	Tile	75.92 (4.99)	73.78 (3.39)	69.89 (4.28)	78.70 (2.84)	74.54 (2.70)	69.63 (3.91)
	DI (Ours)	80.77 (5.48)	74.54 (5.38)	70.31 (6.85)	80.73 (4.05)	75.70 (4.04)	71.41 (4.46)
	Δ	+4.85	+0.76	+0.42	+2.03	+1.16	+1.78
Oral2026	Tile	85.92 (4.96)	85.30 (7.65)	71.74 (8.80)	87.62 (3.40)	84.82 (1.27)	71.63 (6.63)
	DI (Ours)	88.65 (2.75)	86.18 (2.40)	77.14 (5.57)	88.76 (3.34)	84.60 (2.40)	74.38 (5.66)
	Δ	+2.73	+0.88	+5.40	+1.14	-0.22	+2.75

the highest performance in **33 out of 36** backbone, dataset, and metric combinations, while obtaining the second-highest result in one additional setting (DINOv3 on *Oral2026* macro-F1: 72.47 versus 72.57 for DSMIL). Detailed mean with standard deviation results are provided in Appendix A.7.

Performance improvements on *Cervical* are comparatively gradual across all feature backbones. Diagnostic cellular morphology in cervical smears is relatively prominent at the tile level, reducing representation dilution. Furthermore, Smart-CCS and UniCAS incorporate cervical cytopathology during pretraining, establishing strong baseline representations that limit additional gains from instance decomposition.

The strongest performance gains occur under Smart-CCS features, where DI-MIL with ACMIL achieves the highest mean performance across all 12 dataset and metric settings. Minor exceptions occur in classification accuracy on *Oral2026*: MambaMIL achieves 82.90 ACC using DINOv3

Table 2: Comparison with existing MIL methods across four cytopathology datasets and three foundation-model backbones. All methods within each backbone use identical features, data splits, and training protocols. **DI-ACMIL** denotes ACMIL operating on our proposed decomposed instances. Best results are shown in **bold**, and second-best results are underlined.

Method	Cervical			Serous Fluid			Respiratory			Oral2026		
	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
DINOv3												
ABMIL	91.40	84.56	84.40	85.37	78.76	78.70	79.38	74.73	70.64	88.06	79.84	70.93
ACMIL	93.83	87.22	87.01	86.11	79.13	79.11	81.05	75.98	72.11	88.37	80.77	71.45
DTFD	92.29	86.00	85.66	85.13	78.06	77.99	75.77	71.57	66.90	78.98	81.50	58.29
ASMIL	92.57	87.02	86.89	84.31	77.40	77.32	73.11	74.54	69.95	85.02	80.65	58.73
DSMIL	90.00	83.84	83.59	83.06	76.52	76.46	74.42	72.43	68.16	85.60	82.67	**72.57**
TransMIL	91.32	83.44	83.21	85.03	77.40	77.35	75.79	71.47	66.50	79.52	72.53	63.87
MambaMIL	90.94	86.71	86.48	86.00	79.24	79.19	80.38	74.06	70.34	83.22	**82.90**	63.85
HMIL	91.93	84.15	83.84	84.78	77.14	77.03	79.19	74.16	70.28	82.16	81.51	65.58
PTCMIL	91.86	84.66	84.44	85.82	78.50	78.44	79.64	74.16	68.46	86.14	80.60	64.33
DI-ACMIL (Ours)	**93.86**	**88.24**	**88.06**	**86.79**	**80.09**	**80.06**	**82.01**	**76.75**	**72.67**	**89.36**	79.91	72.47
Smart-CCS												
ABMIL	94.38	87.42	87.28	84.22	76.52	76.49	79.36	74.16	70.04	83.91	83.18	67.18
ACMIL	94.31	86.82	86.61	85.04	77.51	77.43	81.21	76.17	71.82	88.20	81.73	72.26
DTFD	92.37	86.71	86.55	83.91	76.30	76.28	76.38	71.37	67.48	82.19	83.16	51.88
ASMIL	91.20	86.61	86.48	78.77	74.35	74.32	72.91	73.39	68.93	78.26	81.79	65.19
DSMIL	93.90	87.22	87.10	81.88	73.76	73.68	75.55	70.60	65.80	84.65	80.92	70.49
TransMIL	93.30	87.12	86.95	83.17	75.45	75.36	75.85	70.99	66.04	86.54	76.92	68.15
MambaMIL	92.92	87.74	87.65	83.00	75.52	75.46	79.16	74.07	70.11	84.32	83.61	67.57
HMIL	93.65	87.12	87.04	83.39	75.52	75.44	78.06	73.39	69.91	84.71	84.05	73.48
PTCMIL	94.61	87.02	86.82	84.87	76.59	76.46	79.13	73.39	69.00	87.79	84.08	73.75
DI-ACMIL (Ours)	**95.89**	**89.98**	**89.87**	**88.33**	**80.64**	**80.60**	**88.85**	**82.32**	**80.08**	**89.13**	**84.52**	**75.69**
UniCAS												
ABMIL	93.68	88.66	88.54	84.04	76.85	76.81	75.92	73.78	69.89	85.92	85.30	71.74
ACMIL	95.42	89.06	88.84	84.45	76.52	76.48	78.70	74.54	69.63	87.62	84.82	71.63
DTFD	92.66	86.30	86.09	82.76	74.97	74.93	70.90	68.30	63.07	85.06	**86.10**	72.42
ASMIL	90.58	87.43	87.30	80.63	74.31	74.28	72.07	72.14	67.04	85.73	82.05	65.13
DSMIL	92.17	86.20	85.98	80.31	73.17	73.08	71.99	71.09	64.88	82.39	83.39	69.75
TransMIL	92.06	84.97	84.78	80.18	73.17	73.12	70.33	68.11	60.96	86.52	84.25	64.57
MambaMIL	93.34	86.51	86.31	83.12	76.22	76.15	76.02	71.18	65.98	83.22	82.20	70.03
HMIL	93.55	87.02	86.91	82.04	75.08	74.95	74.07	71.75	68.71	85.21	82.80	68.86
PTCMIL	94.68	87.73	87.54	83.68	74.05	73.83	76.25	72.14	65.70	86.31	83.71	70.24
DI-ACMIL (Ours)	**95.79**	**89.57**	**89.39**	**86.06**	**78.43**	**78.39**	**80.73**	**75.70**	**71.41**	**88.76**	84.60	**74.38**

compared to 79.91 for DI-MIL with ACMIL, whereas DTFD achieves 86.10 ACC using UniCAS compared to 84.60 for DI-MIL with ACMIL. These results demonstrate that while DI-MIL does not lead every individual metric on smaller imbalanced cohorts, it maintains consistent cross-backbone competitiveness.

4.4 Ablation Studies

We use the ablation studies to separate three potential sources of improvement. First, we vary the number of decomposed instances k to assess granularity sensitivity. Second, with k fixed, we compare feature-based clustering against random and spatial grouping to isolate the effect of the grouping criterion. Lastly, we compare post-encoding decomposition with direct smaller tiling. The full ablation suite is evaluated on both *Oral2026* and *Respiratory*.

4.4.1 Number of decomposed instances

We first vary $k \in \{1, 2, 3, 4, 5\}$ on *Oral2026* and *Respiratory* using Smart-CCS with ACMIL while holding all other settings fixed. As shown in Table 3, $k = 4$ provides the most balanced performance across the two datasets: it achieves the highest AUC and macro-F1 on *Oral2026* and the highest ACC and macro-F1 on *Respiratory*, while $k = 5$ yields the highest *Oral2026* ACC and *Respiratory* AUC. The non-monotonic trend indicates that simply increasing the number of instances is not sufficient. We therefore use $k = 4$ as a practical fixed default rather than claiming a universal optimum. Importantly, the $k = 1$ baseline serves as a control that isolates the feature representation change

from the multi-instance decomposition effect by collapsing dense tokens into a single instance. As shown in Appendix A.4, $k = 4$ consistently outperforms both the original [CLS] baseline and this $k = 1$ control across all datasets.

Table 3: Configuration sweep over the number of decomposed instances k on *Oral2026* and *Respiratory* using Smart-CCS and ACMIL. All configurations use identical data splits and downstream training settings within each dataset and differ only in k. Results are reported as mean with standard deviation in parentheses.

k	Oral2026			Respiratory		
	AUC	ACC	F1	AUC	ACC	F1
1	88.68 (3.93)	83.98 (3.03)	75.19 (4.25)	84.08 (2.16)	78.10 (2.97)	74.51 (3.34)
2	88.86 (4.22)	84.28 (4.43)	75.17 (6.72)	84.94 (2.85)	77.43 (3.34)	74.01 (4.01)
3	88.79 (5.35)	84.00 (6.63)	75.57 (9.55)	87.00 (2.71)	79.25 (3.50)	76.08 (4.72)
4	**89.13** (4.86)	84.52 (5.25)	**75.69** (8.27)	88.85 (2.07)	**82.32** (3.10)	**80.08** (3.48)
5	87.68 (6.23)	**85.06** (5.35)	74.81 (11.59)	**89.32** (2.35)	82.03 (4.64)	79.18 (6.05)

4.4.2 Alternative dense-token grouping strategies

Having fixed $k = 4$, we next ask whether the criterion used to form the four instances matters. Random grouping removes both semantic and spatial structure, whereas spatial 2×2 grouping preserves local position but ignores feature similarity. As shown in Table 4, feature-based DI achieves the highest mean AUC, ACC, and macro-F1 on both *Oral2026* and *Respiratory*. The consistent advantage over both controls indicates that the gain is not explained merely by expanding each tile into four arbitrary instances. Detailed formulations of the two controls are provided in Appendix A.3.

Table 4: Comparison of alternative dense-token grouping strategies on *Oral2026* and *Respiratory* using Smart-CCS and ACMIL. All strategies construct $k = 4$ instances from the same dense-token representations and use identical downstream training settings. Results are reported as mean with standard deviation in parentheses.

Strategy	Oral2026			Respiratory		
	AUC	ACC	F1	AUC	ACC	F1
Random token grouping ($k = 4$)	88.07 (4.41)	83.75 (3.50)	74.79 (5.68)	83.28 (2.42)	76.95 (2.78)	73.60 (3.85)
Spatial 2×2 grouping	86.83 (5.56)	83.38 (4.88)	73.95 (7.29)	86.11 (2.22)	80.11 (1.93)	77.15 (2.61)
DI ($k = 4$)	**89.13** (4.86)	**84.52** (5.25)	**75.69** (8.27)	**88.85** (2.07)	**82.32** (3.10)	**80.08** (3.48)

4.4.3 Decomposed Instances versus Direct Fine-Grid Tiling

The previous comparison shows that post-encoding grouping matters even when the number of instances is fixed. We next test whether comparable gains can instead be obtained by increasing instance granularity before encoding through smaller source tiles. Table 5 compares DI-MIL with progressively finer tiling on *Oral2026* and *Respiratory*. On *Oral2026*, smaller tiling produces non-monotonic changes across tile sizes; on *Respiratory*, smaller tiles progressively improve performance over the native 1200-pixel baseline. Nevertheless, DI-MIL achieves the highest mean AUC, ACC, and macro-F1 on both datasets while retaining the original source-tile encoding context. Thus, its gains cannot be attributed solely to a larger number of MIL instances.

Additionally, we analyze the storage cost which serves as a major practical bottleneck in cytopathology pipelines due to high-resolution tile storage and slide-level I/O. As illustrated for *Oral2026* in Figure 2, direct smaller tiling increases the number of extracted source images by up to $43.3\times$, whereas DI-MIL preserves the original $1\times$ image extraction and encoder cost while extending bag length from N to $4N$. Absolute feature-storage measurements and cost analysis by computational time are further provided in Appendix A.8 A.9.

5 Discussion and Limitations

Conventional WSI MIL operationally treats each encoded tile as an indivisible unit of slide-level aggregation. This formulation is particularly restrictive in cytopathology, where a single microscopic field may contain sparse diagnostic cells intermixed with abundant benign and non-diagnostic content. DI-MIL addresses this fundamental domain mismatch by decoupling visual encoding context from slide aggregation granularity, enabling attention-based aggregators such as ABMIL and

Table 5: Comparison of different tile-size strategies with our decomposed-instance (DI) strategy on *Oral2026* and *Respiratory*. Tile n stands for using pixel tile size n without DI. All methods use Smart-CCS features, ACMIL, and the same downstream training protocol; DI uses the main-experiment setting $k = 4$ with the dataset-specific source tile size (2048 on *Oral2026* and 1200 on *Respiratory*). Results are reported as mean with standard deviation in parentheses.

Strategy	Oral2026			Respiratory		
	AUC	ACC	F1	AUC	ACC	F1
Tile 2048 w/o. DI	88.20 (5.44)	81.73 (6.14)	72.26 (4.25)	–	–	–
Tile 1200 w/o. DI	88.95 (5.06)	84.26 (4.78)	73.89 (4.73)	81.21 (3.45)	76.17 (3.19)	71.83 (4.47)
Tile 600 w/o. DI	88.73 (3.83)	84.24 (5.83)	74.40 (4.75)	83.81 (3.92)	78.10 (3.39)	74.77 (3.22)
Tile 300 w/o. DI	86.20 (3.34)	84.10 (4.88)	74.35 (4.84)	85.49 (2.11)	78.29 (2.93)	75.07 (3.78)
DI ($k = 4$)	**89.13** (4.86)	**84.52** (5.25)	**75.69** (8.27)	**88.85** (2.07)	**82.32** (3.10)	**80.08** (3.48)



Figure 2: **Performance and computational cost comparison between source tile size reduction and DI-MIL on *Oral2026* using Smart-CCS with ACMIL.** Left: AUC and macro-F1 versus the relative number of image tiles. Right: Relative tile count versus MIL feature-instance count.

ACMIL to assign independent importance weights to localized, discriminative cellular patterns. Qualitative inspection of high-attention instances is consistent with this intended granularity effect; examples are provided in Appendix A.5.

On the other hand, DI-MIL has two main limitations. First, its fixed k and non-differentiable clustering do not adapt to variations in tile cellularity or spatial heterogeneity. Second, it treats decomposed instances as an order-agnostic set and therefore does not explicitly model spatial relationships among intra-tile clusters. Potential extensions include adaptive or differentiable decomposition, spatially aware grouping, and task-guided interactions among decomposed instances.

6 Conclusion

We revisited the conventional paradigm in weakly supervised cytopathology that couples tile-level encoding context to MIL aggregation granularity. By decomposing dense tokens from foundation models within each tile into independently weighted instances post-encoding, DI-MIL mitigates the intra-tile evidence dilution that can arise when heterogeneous cytopathology tiles are compressed into single global representations. Extensive evaluations across diverse cytopathology cohorts, foundation backbones, and aggregators demonstrate that DI-MIL consistently outperforms both single-embedding baselines and direct fine-grid tiling. These results demonstrate that the tile-level instance is not always the most effective choice for cytopathology MIL and establish instance construction as a meaningful direction for future study.

AI USE STATEMENT

Generative AI tools were used during preparation of this manuscript to assist with literature search and organization, consistency checking and interpretation of already-computed experimental outputs, and drafting and editing of the paper. The experimental measurements reported in the tables and figures were supplied from the authors' experimental pipeline rather than generated by the language model used for manuscript preparation. The authors reviewed the AI-assisted prose, recomputed the aggregate numerical summaries used in the text from the supplied results, and checked cited claims against the referenced literature. The authors take responsibility for the final content of the work, including text and claims produced with generative-AI assistance.

ETHICS STATEMENT

The private cytopathology cohorts were retrospectively collected and processed under the applicable institutional ethics and data-governance requirements. The study protocol was reviewed and approved by the relevant institutional ethics committee, and the requirement for informed consent was waived because of the retrospective nature of the study. All WSIs and associated labels were de-identified before use. The study involved retrospective model-development experiments only and did not include any prospective clinical intervention. The public *Oral2026* dataset was used in accordance with its released terms. The proposed models are intended for research use and have not been validated for independent clinical deployment.

REPRODUCIBILITY STATEMENT

Section 3 specifies the instance-construction operation and its integration with the evaluated attention-based MIL backends; Section 4.1 defines the controlled comparison axes; and Appendix A.1 reports cohort sizes, tile sizes, cross-validation protocols, split seeds, core training hyperparameters, and the default decomposition configuration. The appendix also provides an algorithmic summary, fold-wise statistics for representative baselines, and the cost accounting used in the tile-size ablation. All paired comparisons use the same underlying features, data splits, and downstream training protocol except for the explicitly varied instance-construction factor. Exact foundation-model checkpoint identifiers, feature-extraction preprocessing conventions, and implementation-level clustering details will accompany the anonymized code/configuration release.

REFERENCES

Mohammad Mahmudul Alam and Mohammad Tariqul Islam. Machine learning approach of automatic identification and counting of blood cells. *Healthcare Technology Letters*, 6(4):103–108, 2019.

Alexandre Bouyssoux, Riadh Fezzani, and Jean-Christophe Olivo-Marin. Cell instance segmentation dataset, feb 4 2022. URL https://zenodo.org/record/5938892.

Joshua Butke, Tatjana Frick, Florian Roghmann, Samir F. El-Mashtoly, Klaus Gerwert, and Axel Mosig. End-to-end multiple instance learning for whole-slide cytopathology of urothelial carcinoma. In *Proceedings of the MICCAI Workshop on Computational Pathology*, volume 156 of *Proceedings of Machine Learning Research*, pp. 57–68. PMLR, 2021.

Mauro Giuseppe Camporeale, Giovanni Dimauro, Matteo Gelardi, Giorgia Iacobellis, Mattia Sebastiano Ladisa, Sergio Latrofa, and Nunzia Lomonte. Nasal mucosa cell dataset (nmcd), 2024. URL https://dx.doi.org/10.21227/0erx-zn98.

Pablo Cuña Cabrera. Apcdata cervical cytology cells, jun 20 2024. URL https://data.mendeley.com/datasets/ytd568rh3p/1.

David Dov, Shahar Z. Kovalsky, Serge Assaad, Jonathan Cohen, Danielle Elliott Range, Avani A. Pendse, Ricardo Henao, and Lawrence Carin. Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images. *Medical Image Analysis*, 67: 101814, 2021. doi: 10.1016/j.media.2020.101814.

Lu Gan, Xi Li, and Xichun Wang. A curated and re-annotated peripheral blood cell dataset integrating four public resources, 2025. URL https://arxiv.org/abs/2407.13214.

Anubha Gupta, Shiv Gehlot, Shubham Goswami, Sachin Motwani, Ritu Gupta, Álvaro García Faura, Dejan Štepec, Tomaž Martinčič, Reza Azad, Dorit Merhof, Afshin Bozorgpour, Babak Azad, Alaa Sulaiman, Deepanshu Pandey, Pradyumna Gupta, Sumit Bhattacharya, Aman Sinha, Rohit Agarwal, Xinyun Qiu, Yucheng Zhang, Ming Fan, Yoonbeom Park, Daehong Lee, Joon Sik Park, Kwangyeol Lee, and Jaehyung Ye. Segpc-2021: A challenge & dataset on segmentation of multiple myeloma plasma cells from microscopic images. *Medical Image Analysis*, 83:102677, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102677. URL https://www.sciencedirect.com/science/article/pii/S136184152200305X.

Balazs Harangi, Gergo Bogacsovics, Janos Toth, Ilona Kovacs, Erzsebet Dani, and Andras Hajdu. Pixel-wise segmentation of cells in digitized Pap smear images. *Scientific Data*, 11(1), jul 6 2024. ISSN 2052-4463. doi: 10.1038/s41597-024-03566-9. URL http://dx.doi.org/10.1038/s41597-024-03566-9.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 2018.

Garima Jain, Abhijeet Patil, Poonam Goel, Sanghamitra Pati, Amit Sethi, Gururaj Malekar, Nilesh Kowe, Nishi Halduniya, Jatin Kashyap, Divyajeet Rout, Sharat Kumar, Hitesh, Heena Tabassum, Rupinder Singh Dhaliwal, Meeta Singh, Ravi Meher, Sucheta Devi Khuraijam, Sushma Khuraijam, Sharmila Laishram, Simmi Kharb, Sunita Singh, K. Swaminadtan, Ranjana Solanki, Deepika Hemranjani, Shashank Nath Singh, Uma Handa, Manveen Kaur, Surinder Singhal, Shivani Kalhan, Rakesh Kumar Gupta, S. Ravi, D. Pavithra, Sunil Kumar Mahto, Arvind Kumar, Deepali Tirkey, Saurav Banerjee, and L. Sreelakshmi. A multi-center annotated oral cytology dataset for ai-assisted early detection of oral squamous cell carcinoma. *Scientific Data*, 2026. doi: 10.1038/s41597-026-07932-7. URL https://doi.org/10.1038/s41597-026-07932-7.

Jan Jantzen, Jonas Norup, Georgios Dounias, and Beth Bjerregaard. Pap-smear benchmark data for pattern classification. In *Proc. NiSIS 2005*, pp. 1–9. NiSIS, 2005. Nature inspired Smart Information Systems : EU co-ordination action, Nisis 2005.

Hao Jiang, Rushan Zhang, Yanning Zhou, Yumeng Wang, and Hao Chen. Donet: Deep de-overlapping network for cytology instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15641–15650, 2023.

Hao Jiang, Cheng Jin, Huangjing Lin, Yanning Zhou, Xi Wang, Jiabo Ma, Li Ding, Jun Hou, Runsheng Liu, Zhizhong Chai, Luyang Luo, Huijuan Shi, Yinling Qian, Qiong Wang, Changzhong Li, Anjia Han, Ronald Cheong Kin Chan, and Hao Chen. Generalizable cervical cancer screening via large-scale pretraining and test-time adaptation, 2025.

Haotian Jiang, Jiangdong Cai, Zhenrong Shen, Mengjie Xu, Manman Fei, Haolin Huang, Xinyu Wang, Rui Bi, Dinggang Shen, Lichi Zhang, and Qian Wang. UniCAS: A foundation model for cervical cytology screening. *Cell Reports Medicine*, 7(1):102570, 2026. doi: 10.1016/j.xcrm.2025.102570.

Cheng Jin, Luyang Luo, Huangjing Lin, Jun Hou, and Hao Chen. Hmil: Hierarchical multi-instance learning for fine-grained whole slide image classification. *IEEE Transactions on Medical Imaging*, 44(4):1796–1808, 2025. doi: 10.1109/TMI.2024.3520602.

Jue Wang. Artificial Intelligence Enables Precision Diagnosis of Cervical Cytology Grades and Cervical Cancer, mar 18 2024. URL https://zenodo.org/doi/10.5281/zenodo.10828394.

Dongwoo Kim, Jongwon Lee, Minsoo Jung, Kwangil Yim, Gisu Hwang, Hongjun Yoon, Daeky Jeong, Won June Cho, Mohammad Rizwan Alam, Gyungyub Gong, Nam Hoon Cho, Chong Woo Yoo, Yosep Chong, and Kyung Jin Seo. Whole slide image-level classification of malignant effusion cytology using clustering-constrained attention multiple instance learning. *Lung Cancer*, 204:108552, 2025. doi: 10.1016/j.lungcan.2025.108552.

Zahra Mousavi Kouzehkanan, Sepehr Saghari, Sajad Tavakoli, Peyman Rostami, Mohammadjavad Abaszadeh, Farzaneh Mirzadeh, Esmaeil Shahabi Satlsar, Maryam Gheidishahran, Fatemeh Gorgi, Saeed Mohammadi, and Reshad Hosseini. A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm. *Scientific Reports*, 12(1), jan 21 2022. ISSN 2045-2322. doi: 10.1038/s41598-021-04426-x. URL http://dx.doi.org/10.1038/s41598-021-04426-x.

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

Honglin Li, Zhongyi Shui, Yunlong Zhang, Chenglu Zhu, and Lin Yang. PathVQ: Reforming computational pathology foundation model for whole slide image analysis via vector quantization. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Paul Mooney. Blood cell images, Apr 2018. URL https://www.kaggle.com/datasets/paultimothymooney/blood-cells.

Rupam Mukherjee, Rajkumar Daniel, Soujanya Hazra, Shirin Dasgupta, and Subhamoy Mandal. Raa-mil: A novel framework for classification of oral cytology, 2025. URL https://arxiv.org/abs/2511.12269.

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*, volume 34, 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025.

Feng Su, Yu Sun, Yajie Hu, Peijiang Yuan, Xinyu Wang, Qian Wang, Jianmin Li, and Jia-Fu Ji. Development and validation of a deep learning system for ascites cytopathology interpretation. *Gastric Cancer*, 23(6):1041–1050, jun 4 2020. ISSN 1436-3291. doi: 10.1007/s10120-020-01093-1. URL http://dx.doi.org/10.1007/s10120-020-01093-1.

André Victória Matias, Allan Cerentini, Luiz Antonio Buschetto Macarini, João Gustavo Atkinson Amorim, Felipe Perozzo Daltoé, and Aldo von Wangenheim. Segmentation, detection and classification of cell nuclei on oral cytology samples stained with papanicolaou. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 53–58, 2020. doi: 10.1109/CBMS49503.2020.00018.

Thomas Wimmer, Prune Truong, Marie-Julie Rakotosaona, Michael Oechsle, Federico Tombari, Bernt Schiele, and Jan Eric Lenssen. Anyup: Universal feature upsampling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Shu Yang, Yihui Wang, and Hao Chen. MambaMIL: Enhancing long sequence modeling with sequence reordering in computational pathology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15004, pp. 296–306. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72083-3_28.

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 *International Conference on Learning Representations*, 2026.

Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

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

Jianwei Zhang. Data for: Binary Tree-like Network with Two-path Fusion Attention Feature for Cervical Cell Nucleus Segmentation, 2019. URL https://data.mendeley.com/datasets/jks43dkjj7/1.

Xuan Zhang, Jianxin Ji, Qi Zhang, Xiaohan Zheng, Kaiyuan Ge, Menglei Hua, Lei Cao, and Liuying Wang. A large annotated cervical cytology images dataset for AI models to aid cervical cancer screening. *Scientific Data*, 12(1), jan 7 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-04374-5. URL http://dx.doi.org/10.1038/s41597-025-04374-5.

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

Beidi Zhao, Wenlong Deng, Zi Han Henry Li, Chen Zhou, Zuhua Gao, Gang Wang, and Xiaoxiao Li. LESS: Label-efficient multi-scale learning for cytological whole slide image screening. *Medical Image Analysis*, 94:103109, 2024. doi: 10.1016/j.media.2024.103109.

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*, 111:104013, 2026. doi: 10.1016/j.media.2026.104013.

Jing Zhao, Yong jun He, Shu-Hang Zhou, Jian Qin, and Yi ning Xie. Cnseg: A dataset for cervical nuclear segmentation. *Computer Methods and Programs in Biomedicine*, 241:107732, 2023. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2023.107732. URL https://www.sciencedirect.com/science/article/pii/S016926072300398X.

Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018. doi: https://doi.org/10.1016/j.micron.2018.01.010. URL https://www.sciencedirect.com/science/article/pii/S0968432817303037.

A Appendix

A.1 Datasets, Splits, and Training Protocol

Table 6 summarizes the cohort and evaluation settings. *Cervical*, *Serous Fluid*, and *Respiratory* are private cytopathology cohorts scanned at $0.5\,\mu$m/pixel. We extract non-overlapping 1200×1200-pixel tiles using a stride of 1200 pixels after Otsu-based background filtering.

The same folds are used by all methods, and no tiles from the same WSI appear in different folds. Reported values are the mean and standard deviation across the corresponding cross-validation folds.

The token grid is determined by the backbone input resolution after preprocessing and is distinct from the native WSI source-tile size reported in Table 6.

Table 6: Dataset and evaluation settings used in the experiments. Positive rate denotes the positive/cancerous class fraction. Avg. tiles / slide denotes the mean number of retained source tiles per WSI after background filtering.

Dataset	Specimen	Slides	Task	Positive rate	Tile size	CV	Avg. tiles / slide
Cervical	Cervical	978	Negative vs. Positive	53.3%	1200 px	10-fold	1940
Serous Fluid	Serous fluid	2,717	Negative vs. Positive	51.5%	1200 px	10-fold	1572
Respiratory	Respiratory	1,041	Negative vs. Positive	66.3%	1200 px	10-fold	1141
Oral2026	Oral	363	Non-cancerous vs. Cancerous	17.6%	2048 px	5-fold	2579

Table 7: Details of the backbones used in the experiments.

Backbone	Encoder patch size (px)	Token grid	Dense tokens / encoded tile
Smart-CCS	14	16×16	256
DINOv3 (vit-l16)	16	14×14	196
UniCAS	16	14×14	196

These statistics also clarify the scale of the instance-construction step. A conventional bag contains N tile embeddings, whereas DI-MIL with $k = 4$ contains $4N$ instances. Directly exposing all dense patch tokens to the MIL aggregator would instead produce $196N$ or $256N$ instances depending on the backbone, motivating the intermediate clustering step.

Across datasets, backbones, and the paired MIL experiments, training is standardized wherever the backend permits. Table 8 lists the common settings and the backend-specific choices. *Oral2026* uses class-weighted cross-entropy because of its smaller and more imbalanced cohort.

Table 8: Core training and decomposition settings. Backend-specific optimizer/objective choices follow the corresponding official MIL formulation.

Setting	Value
Learning rate	2×10^{-4}
Training seed	1024
Optimizer	Adam (ABMIL); AdamW (ACMIL)
Objective	CE (ABMIL); sub-token CE + slide CE + attention-diversity loss (ACMIL)
Default decomposition	$k = 4$; 8 clustering iterations
Early stopping / maximum epochs	Enabled / 100
Feature dimension / classes	1024 / 2
Input normalization	LayerNorm enabled
Batch size	1 WSI (one bag)
Weight decay / dropout	10^{-5} / 0.25
Metrics	AUC, ACC, macro-F1

A.2 Analysis of Fine-Grained Dense-Token Representations

The main experiments show that DI-MIL yields larger improvements with the cytopathology-specific Smart-CCS and UniCAS backbones than with the general-purpose DINOv3 backbone. To further examine this backbone-dependent behavior, we analyze whether their dense-token representations differ in their ability to preserve localized cellular evidence, using both quantitative evaluation and qualitative visualization.

Quantitative analysis. We evaluate the frozen dense-token representations on 18 public datasets with cell-level bounding-box annotations or bounding boxes derived from available instance masks. For each annotated object, we extract a 224×224 crop centered on its ground-truth bounding box and obtain the corresponding dense patch-token embeddings. A token is regarded as positive evidence if the center of its spatial support lies within the annotated bounding box. Using validation data only, we compute positive- and negative-evidence prototypes, denoted by $\boldsymbol{\mu}_+$ and $\boldsymbol{\mu}_-$. For each test token $\mathbf{z}_i$, we define an evidence score as

$$s_i = \sigma\left[\cos(\mathbf{z}_i, \boldsymbol{\mu}_+) - \cos(\mathbf{z}_i, \boldsymbol{\mu}_-)\right], \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Test tokens are ranked by s_i, and we measure how rapidly tokens corresponding to annotated cellular evidence are recovered as progressively larger fractions of the ranked tokens are included. We summarize this behavior using the random-corrected insertion AUC gain

$$G = \mathrm{AUC}(I) - \mathrm{AUC}(I_{\mathrm{rnd}}), \tag{8}$$

where I denotes the insertion-recall curve and I_{rnd} denotes the corresponding curve under a deterministic random token ordering. A larger G indicates that annotated cellular evidence is more readily identifiable from the frozen dense-token representation. Across the 18 datasets, Smart-CCS achieves the highest equal-weight mean insertion gain ($G = 0.2254$), followed by UniCAS (0.2173) and DINOv3 (0.1996). Smart-CCS and UniCAS outperform DINOv3 on 14/18 and 15/18 datasets, respectively, with one-sided paired Wilcoxon p-values of 0.0014 and 0.00042. As shown in Figure 3, both cytopathology-specific backbones also achieve higher mean insertion gains than DINOv3 across token budgets from 5% to 50%. Detailed dataset-level results are provided in Table 9.

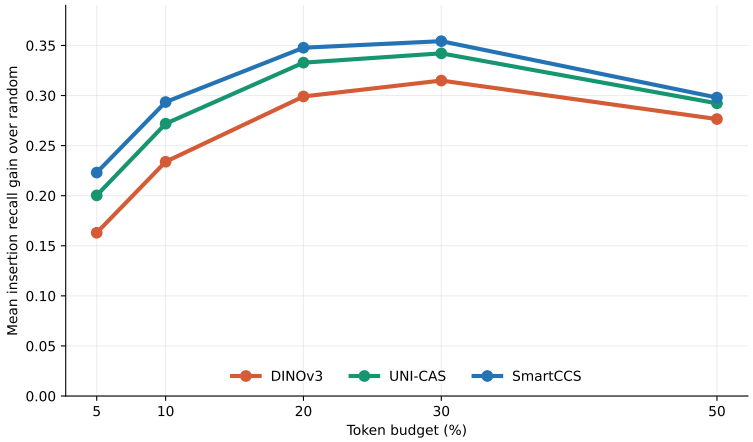


Figure 3: Mean random-corrected insertion-recall gain across 18 public cell-level datasets as a function of retained token budget. Higher values indicate that annotated cellular evidence is recovered more readily from the frozen dense-token representation.

Table 9: Dataset-level random-corrected insertion AUC gain G for the three frozen backbones. Higher values indicate that annotated cellular evidence is recovered more readily from the dense-token representation. $S - D$ and $U - D$ denote the paired differences of Smart-CCS and UniCAS relative to DINOv3, respectively.

Dataset	Smart-CCS	UniCAS	DINOv3	$S - D$	$U - D$	Test objects
APACS23 (Harangi et al., 2024)	0.1023	**0.1045**	0.0983	+0.0040	+0.0062	8,671
APC (Cuña Cabrera, 2024)	**0.2263**	0.2112	0.1569	+0.0694	+0.0542	281
Ascites2020 (Su et al., 2020)	**0.4293**	0.4213	0.3499	+0.0794	+0.0714	581
BCCD (Mooney, 2018)	0.1279	0.1279	**0.1286**	-0.0007	-0.0007	71
BTTFA (Zhang, 2019)	**0.4551**	0.4015	0.3914	+0.0637	+0.0101	978
CBC (Alam & Islam, 2019)	0.0532	0.0485	**0.0548**	-0.0016	-0.0063	61
CellDet (Zhang et al., 2025)	0.1769	**0.1816**	0.1550	+0.0219	+0.0266	3,104
CISD (Bouyssoux et al., 2022)	**0.0807**	0.0682	0.0636	+0.0171	+0.0046	921
CNSeg (Zhao et al., 2023)	**0.4028**	0.3752	0.3507	+0.0521	+0.0246	10,649
CPS (Jiang et al., 2023)	0.2667	**0.2815**	0.2241	+0.0426	+0.0573	1,209
Herlev (Jantzen et al., 2005)	0.1298	**0.1325**	0.1312	-0.0014	+0.0014	399
Jiangxi (Zheng et al., 2018)	0.3180	0.3267	**0.3270**	-0.0091	-0.0004	134
NMCD (Camporeale et al., 2024)	**0.2424**	0.2054	0.1577	+0.0847	+0.0477	1,057
Oral2021 (Jue Wang, 2024)	**0.3712**	0.3587	0.3559	+0.0153	+0.0028	253
Raabin-WBC (Kouzehkanan et al., 2022)	**0.0917**	0.0891	0.0827	+0.0090	+0.0064	461
SegPC (Gupta et al., 2023)	0.0248	**0.0249**	0.0243	+0.0005	+0.0006	948
TXL-PBC (Gan et al., 2025)	0.1389	**0.1440**	0.1383	+0.0006	+0.0057	155
UFSC-OCPap (Victória Matias et al., 2020)	**0.4187**	0.4084	0.4027	+0.0160	+0.0057	5,081
Equal-weight mean	**0.2254**	0.2173	0.1996	+0.0258	+0.0177	35,014

Qualitative analysis. For qualitative visualization, the dense token feature maps are upsampled to the input-image resolution using AnyUp (Wimmer et al., 2026). This upsampling is used only for qualitative visualization and is not involved in token clustering, DI construction, or downstream MIL training.

Figure 4 provides qualitative examples of the dense token representations from the three evaluated backbones and the corresponding intra-tile clustering used to construct decomposed instances. The examples suggest that the cytopathology-specific foundation models, Smart-CCS and UniCAS, exhibit more distinct dense-token representations of abnormal cytomorphological patterns than DINOv3.

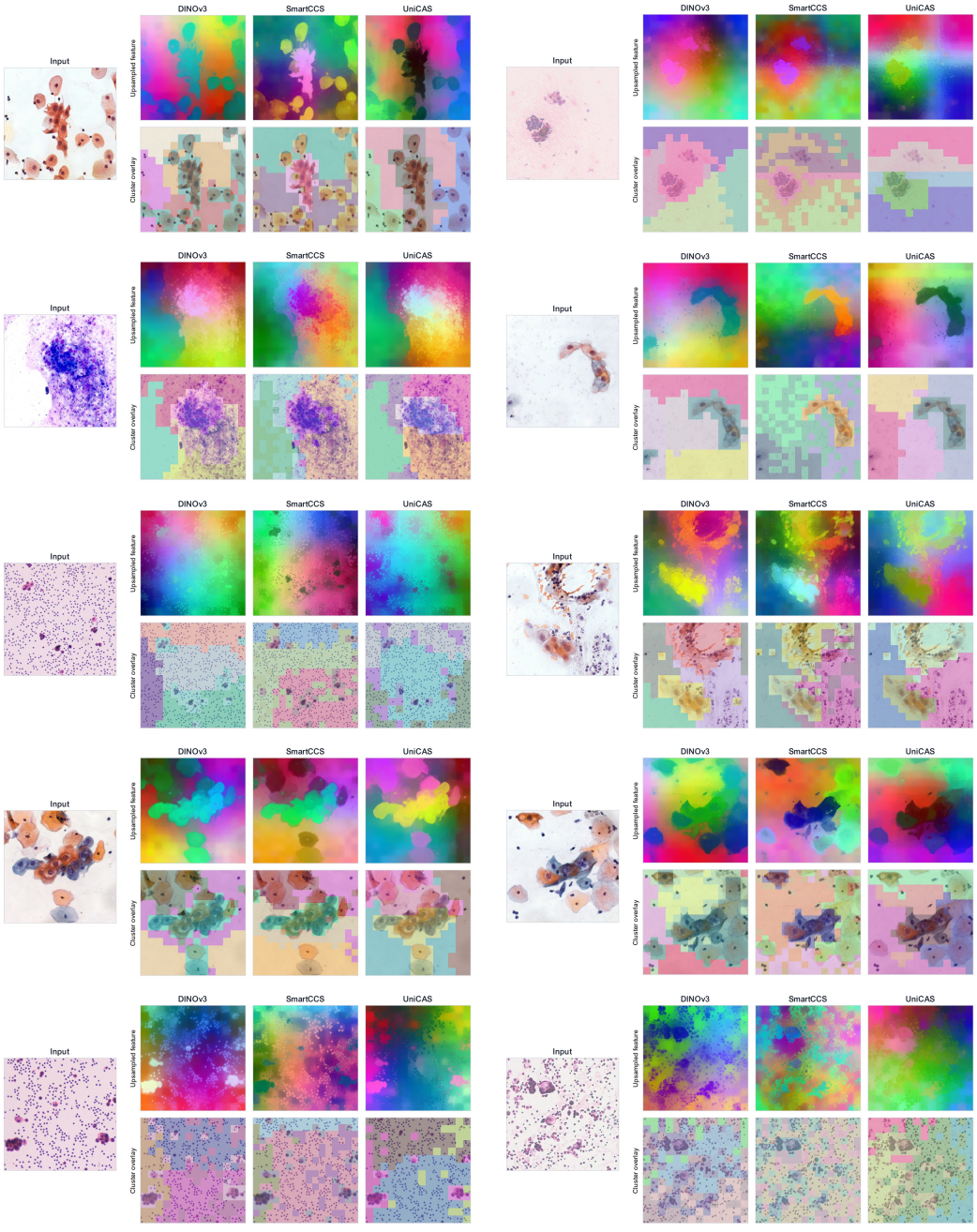


Figure 4: Visualization of dense token representations and decomposed instances. Patch tokens within the same color group form one DI. The color difference is positively correlated to feature vector distance in the encoder feature space.

A.3 Alternative Dense-Token Grouping Strategies

To isolate the effect of feature-based token grouping in DI-MIL, we compare the proposed spherical clustering with two alternative strategies that construct the same number of instances from the same dense patch-token representations. Specifically, we consider *random grouping* and *spatial* 2×2 *grouping*. Both alternatives operate independently within each source tile and use $k = 4$, matching the default DI-MIL configuration. They differ from DI-MIL only in how patch tokens are assigned to groups; the subsequent instance pooling and slide-level MIL aggregation remain unchanged.

Let the M patch-token representations extracted from tile p_i be $\{\boldsymbol{h}_{ij}\}_{j=1}^{M}$, following the notation in Section 3.

Random grouping. Random grouping removes both feature similarity and spatial structure from the decomposition procedure. For each tile p_i, we randomly permute the token indices $\{1, \dots, M\}$ and partition the resulting sequence into $k = 4$ disjoint groups of approximately equal size,

$$\{\mathcal{R}_{i1}, \dots, \mathcal{R}_{ik}\} = \mathrm{Split}_k(\mathrm{Permute}(\{1, \dots, M\})), \qquad \bigcup_{\ell=1}^{k} \mathcal{R}_{i\ell} = \{1, \dots, M\}. \tag{9}$$

When M is not exactly divisible by k, group sizes differ by at most one token. Each random group is then converted into one MIL instance using the same pooling operation as DI-MIL,

$$\boldsymbol{d}_{i,\mathrm{rand}}^{\ell} = \mathrm{LN}\left(\frac{1}{|\mathcal{R}_{i\ell}|} \sum_{j \in \mathcal{R}_{i\ell}} \boldsymbol{h}_{ij}\right), \qquad \ell = 1, \dots, 4. \tag{10}$$

Thus, each source tile still contributes four independently weighted MIL instances, but membership in an instance is unrelated to either token semantics or spatial location. This baseline tests whether the performance gain can be explained simply by increasing the number of instances constructed from dense tokens.

Spatial 2×2 grouping. Spatial grouping instead preserves the original spatial organization of the transformer token grid while ignoring feature similarity. Suppose the M patch tokens of tile p_i form an $H \times W$ spatial grid, with token j located at grid coordinate (r_j, c_j). We divide this grid into four non-overlapping spatial quadrants according to the horizontal and vertical midpoints,

$$\mathcal{S}_i^{ab} = \{j \mid \mathbb{I}(r_j \geq H/2) = a, \mathbb{I}(c_j \geq W/2) = b\}, \qquad a, b \in \{0, 1\}, \tag{11}$$

yielding the four groups corresponding to the top-left, top-right, bottom-left, and bottom-right regions of the token grid. Each spatial group is pooled identically to DI-MIL,

$$\boldsymbol{d}_{i,\mathrm{sp}}^{ab} = \mathrm{LN}\left(\frac{1}{|\mathcal{S}_i^{ab}|} \sum_{j \in \mathcal{S}_i^{ab}} \boldsymbol{h}_{ij}\right), \qquad a, b \in \{0, 1\}. \tag{12}$$

Spatial 2×2 grouping therefore also produces exactly four instances per tile, but constrains each instance to a contiguous spatial region. In contrast, DI-MIL groups tokens according to similarity in the encoder representation space and does not require grouped tokens to be spatially adjacent.

For all three strategies, the resulting instances retain the original feature dimensionality and are assembled across tiles into a slide-level bag of size $4N$. Consequently, the encoder inputs, dense-token representations, number of MIL instances, pooling operation, and downstream aggregator are held fixed. The comparison in Section 4.4.2 therefore primarily evaluates the effect of the token-grouping criterion itself: random assignment, fixed spatial partitioning, or feature-based clustering.

A.4 Controlling for Dense-Token Mean Pooling ($k = 1$)

The main Tile–DI comparison changes both the representation exposed to the MIL model and the number of independently weighted instances: the Tile baseline uses the encoder [CLS] embedding, whereas DI is constructed from dense patch-token representations. We therefore use $k = 1$ as a representation control. With $k = 1$, all dense tokens within a tile are mean-pooled into a single

instance, so the bag still contains exactly one instance per source tile. Across all four datasets using Smart-CCS with ACMIL, $k = 4$ outperforms both the [CLS] baseline and the $k = 1$ control on all 12 dataset–metric comparisons. The $k = 1$ representation is not uniformly better than [CLS], indicating that the main improvement cannot be explained solely by replacing the global token with dense-token mean pooling.

Table 10: Comparison of the conventional tile-level [CLS] representation, dense-token mean pooling ($k = 1$), and the default decomposed-instance setting ($k = 4$) across all four datasets. All variants use Smart-CCS, ACMIL, identical dataset-native source tiles, and the same downstream training protocol.

Oral2026			
Method	AUC	ACC	F1
[CLS]	88.20 (5.44)	81.73 (6.14)	72.26 (4.25)
DI ($k = 1$)	88.68 (3.93)	83.98 (3.03)	75.19 (4.25)
DI ($k = 4$)	**89.13** (4.86)	**84.52** (5.25)	**75.69** (8.27)

Cervical			
Method	AUC	ACC	F1
[CLS]	94.31 (1.90)	86.82 (3.36)	86.61 (3.48)
DI ($k = 1$)	94.66 (1.46)	88.65 (2.34)	88.46 (2.46)
DI ($k = 4$)	**95.89** (1.21)	**89.98** (3.36)	**89.87** (3.35)

Serous Fluid			
Method	AUC	ACC	F1
[CLS]	85.04 (2.82)	77.51 (3.20)	77.43 (3.25)
DI ($k = 1$)	84.16 (2.63)	76.44 (2.24)	76.41 (2.25)
DI ($k = 4$)	**88.33** (2.09)	**80.64** (2.62)	**80.60** (2.63)

Respiratory			
Method	AUC	ACC	F1
[CLS]	81.21 (3.45)	76.17 (3.19)	71.83 (4.47)
DI ($k = 1$)	84.08 (2.16)	78.10 (2.97)	74.51 (3.34)
DI ($k = 4$)	**88.85** (2.07)	**82.32** (3.10)	**80.08** (3.48)

A.5 Qualitative Analysis of High-Attention Instances

We further inspect highly attended instances from conventional tile-level MIL and DI-MIL. Figure 5 shows examples across the four datasets. In these selected cases, decomposed instances tend to concentrate on more localized cellular content while excluding portions of surrounding background, staining variation, or artifacts. These examples are intended to illustrate the behavior of the learned attention and do not constitute quantitative localization evaluation.

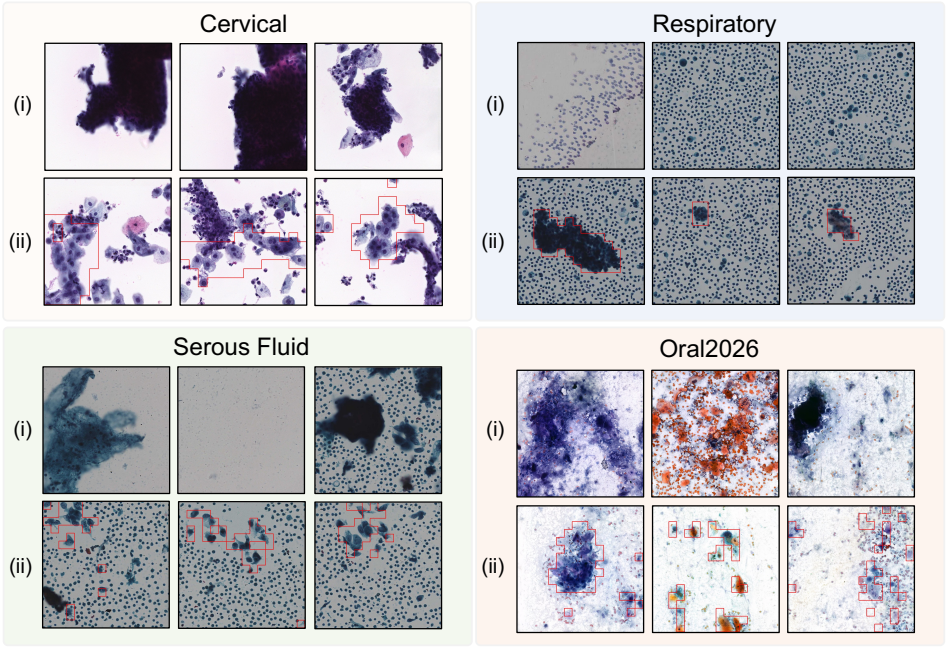


Figure 5: Qualitative comparison of high-attention instances across the four datasets. Row (i) shows conventional tile instances, whereas row (ii) shows decomposed instances, with red contours marking regions assigned to high-attention DIs. These examples are illustrative and do not constitute quantitative localization evaluation.

A.6 Paired Fold-level Statistics for Main Result

To assess whether the mean improvements in Table 1 are stable across matched folds, we additionally report fold-wise DI-minus-Tile differences. These analyses use the fold as the pairing unit and are intended to characterize effect consistency across the fixed cross-validation splits. Because training sets overlap across cross-validation folds, the resulting p-values should be interpreted as descriptive fold-paired evidence rather than as tests based on fully independent experimental replicates.

Table 11: Fold-paired tests of Decomposed instances (DI) versus Tile on AUC. Each comparison uses the same folds, features, splits, and MIL protocol as Table 1. Δ denotes the mean fold-wise difference DI $-$ Tile in percentage points, and values in parentheses are SD_Δ, the standard deviation of the paired differences. The 95% CI is the paired Student's t interval for the mean Δ. p_W is the two-sided exact Wilcoxon signed-rank p-value (zero differences dropped), and p_t is the two-sided paired t-test p-value. n_+ counts folds with $\Delta > 0$. The three institutional cohorts use 10 folds, whereas *Oral2026* uses 5 folds. For *Oral2026*, the smallest attainable two-sided exact Wilcoxon p-value is 0.0625 when all five paired differences are non-zero.

ABMIL					
Dataset	n	Δ (SD_Δ)	95% CI	p_W / p_t	n_+
DINOv3					
Cervical	10	+0.12 (1.66)	$[-1.07, 1.31]$	0.695 / 0.826	6/10
Serous Fluid	10	+0.42 (2.12)	$[-1.09, 1.93]$	0.695 / 0.545	6/10
Respiratory	10	+1.38 (4.61)	$[-1.92, 4.68]$	0.695 / 0.368	5/10
Oral2026	5	-2.34 (1.92)	$[-4.72, 0.04]$	0.062 / 0.053	0/5
Smart-CCS					
Cervical	10	+0.33 (2.22)	$[-1.26, 1.91]$	0.557 / 0.654	6/10
Serous Fluid	10	+3.41 (1.59)	$[2.27, 4.55]$	0.002^{**} / $< 0.001^{***}$	10/10
Respiratory	10	+8.14 (3.56)	$[5.59, 10.68]$	0.002^{**} / $< 0.001^{***}$	10/10
Oral2026	5	+0.26 (2.28)	$[-2.57, 3.08]$	0.812 / 0.813	3/5
UniCAS					
Cervical	10	+0.67 (1.90)	$[-0.69, 2.03]$	0.375 / 0.292	6/10
Serous Fluid	10	+1.09 (2.95)	$[-1.02, 3.20]$	0.232 / 0.274	7/10
Respiratory	10	+4.85 (5.05)	$[1.24, 8.46]$	0.020^{*} / 0.014^{*}	8/10
Oral2026	5	+2.73 (3.94)	$[-2.16, 7.62]$	0.250 / 0.196	3/5

ACMIL					
Dataset	n	Δ (SD_Δ)	95% CI	p_W / p_t	n_+
DINOv3					
Cervical	10	+0.03 (2.74)	$[-1.93, 1.99]$	1.00 / 0.973	5/10
Serous Fluid	10	+0.68 (2.90)	$[-1.39, 2.76]$	0.625 / 0.476	6/10
Respiratory	10	+0.96 (1.97)	$[-0.46, 2.37]$	0.232 / 0.160	5/10
Oral2026	5	+0.99 (2.17)	$[-1.70, 3.68]$	0.438 / 0.366	3/5
Smart-CCS					
Cervical	10	+1.59 (2.72)	$[-0.36, 3.53]$	0.160 / 0.098	6/10
Serous Fluid	10	+3.30 (2.29)	$[1.66, 4.94]$	0.002^{**} / 0.001^{**}	10/10
Respiratory	10	+7.64 (3.70)	$[5.00, 10.29]$	0.004^{**} / $< 0.001^{***}$	9/10
Oral2026	5	+0.93 (3.44)	$[-3.34, 5.21]$	0.812 / 0.577	3/5
UniCAS					
Cervical	10	+0.37 (1.56)	$[-0.74, 1.48]$	0.492 / 0.472	7/10
Serous Fluid	10	+1.61 (1.95)	$[0.21, 3.01]$	0.037^{*} / 0.028^{*}	8/10
Respiratory	10	+2.02 (2.56)	$[0.19, 3.86]$	0.064 / 0.034^{*}	8/10
Oral2026	5	+1.13 (3.95)	$[-3.77, 6.03]$	0.625 / 0.556	4/5

Table 12: Per-fold AUC Δ (DI − Tile, percentage points) for every matched comparison in Table 1. The fold is the pairing unit. Med. $[Q_1, Q_3]$ summarises the fold-wise Δ distribution.

Dataset	MIL	0	1	2	3	4	5	6	7	8	9	Med. $[Q_1, Q_3]$
DINOv3												
Cervical	ABMIL	+0.55	+0.42	+0.44	-1.05	+2.26	-0.36	+2.34	-2.81	-1.79	+1.20	$+0.43\,[-0.87, +1.03]$
	ACMIL	-3.85	-1.90	+4.31	+1.59	-3.22	0.00	+3.14	+2.05	-1.88	+0.06	$+0.03\,[-1.90, +1.94]$
Serous Fluid	ABMIL	+0.49	+3.64	-2.13	+1.03	+0.15	+4.19	-1.73	-1.49	+0.33	-0.27	$+0.24\,[-1.19, +0.89]$
	ACMIL	+3.92	+4.11	-2.16	+3.51	-1.27	+2.07	-4.23	-1.87	+1.11	+1.65	$+1.38\,[-1.72, +3.15]$
Respiratory	ABMIL	-2.58	-2.69	+2.53	-0.29	+9.15	+0.25	+5.92	+7.16	-4.31	-1.30	$-0.02\,[-2.26, +5.07]$
	ACMIL	+2.01	+3.56	-0.68	-0.02	+3.19	-0.95	-1.70	-0.77	+2.90	+2.03	$+1.00\,[-0.75, +2.68]$
Oral2026	ABMIL	-0.14	-1.13	-3.13	-5.13	-2.18	–	–	–	–	–	$-2.18\,[-3.13, -1.13]$
	ACMIL	+0.41	+0.63	+4.77	-0.38	-0.48	–	–	–	–	–	$+0.41\,[-0.38, +0.63]$
Smart-CCS												
Cervical	ABMIL	-0.88	-1.80	+4.22	-0.84	+1.42	+1.34	+1.42	+1.84	-3.70	+0.21	$+0.78\,[-0.87, +1.42]$
	ACMIL	-0.50	-1.80	+0.67	+3.26	+3.43	+3.51	+5.58	+4.28	-0.79	-1.77	$+1.96\,[-0.72, +3.49]$
Serous Fluid	ABMIL	+0.90	+2.79	+3.99	+4.58	+3.09	+2.65	+2.47	+4.54	+6.66	+2.44	$+2.94\,[+2.51, +4.40]$
	ACMIL	+0.70	+2.36	+7.18	+1.13	+2.67	+3.93	+0.29	+4.90	+5.95	+3.87	$+3.27\,[+1.43, +4.66]$
Respiratory	ABMIL	+7.19	+6.36	+13.19	+12.09	+7.56	+12.32	+3.64	+9.86	+5.76	+3.44	$+7.37\,[+5.91, +11.53]$
	ACMIL	+11.76	-1.06	+5.28	+9.67	+7.91	+11.43	+6.83	+9.94	+7.83	+6.87	$+7.87\,[+6.84, +9.87]$
Oral2026	ABMIL	+1.65	-2.63	-1.64	+2.69	+1.21	–	–	–	–	–	$+1.21\,[-1.64, +1.65]$
	ACMIL	-1.38	+5.26	+1.34	-3.46	+2.91	–	–	–	–	–	$+1.34\,[-1.38, +2.91]$
UniCAS												
Cervical	ABMIL	+0.33	+2.51	+0.75	+0.38	+4.22	-0.06	-1.38	+2.49	-1.88	-0.64	$+0.36\,[-0.50, +2.06]$
	ACMIL	+0.59	+1.80	+0.56	+1.25	-2.59	-1.51	+2.17	-0.67	+0.21	+1.88	$+0.57\,[-0.45, +1.66]$
Serous Fluid	ABMIL	+3.63	+0.32	-0.71	+2.94	+2.45	+2.27	-5.09	-1.33	+5.26	+1.13	$+1.70\,[-0.45, +2.82]$
	ACMIL	+2.01	+1.47	-1.25	+3.61	+0.92	+2.23	-1.99	+3.35	+1.89	+3.87	$+1.95\,[+1.06, +3.07]$
Respiratory	ABMIL	+11.35	+1.97	+3.58	+6.79	+5.22	-0.41	-1.78	+8.61	+0.23	+12.96	$+4.40\,[+0.66, +8.16]$
	ACMIL	+1.57	+2.94	+5.05	+1.01	+2.96	-2.11	-2.11	+4.97	+4.06	+1.90	$+2.42\,[+1.15, +3.78]$
Oral2026	ABMIL	0.00	+9.02	-0.45	+4.10	+0.97	–	–	–	–	–	$+0.97\,[+0.00, +4.10]$
	ACMIL	+3.17	+2.51	+0.89	-5.51	+4.61	–	–	–	–	–	$+2.51\,[+0.89, +3.17]$

Table 13: Fold-paired tests of DI versus Tile on ACC and macro-F1, using the same matched folds as Table 1. ACC and macro-F1 are treated as secondary metrics and are reported for completeness; their fold-paired analyses should not be interpreted as independent replications of the AUC analysis. Mean Δ, 95% paired-t CI, and two-sided exact Wilcoxon p_{W} follow Table 11. $^{*}\,p < 0.05$, $^{**}\,p < 0.01$, $^{***}\,p < 0.001$.

Dataset	Metric	ABMIL			ACMIL		
		Mean Δ	95% CI	p_{W}	Mean Δ	95% CI	p_{W}
DINOv3							
Cervical	ACC	+1.13	[−1.64, 3.89]	0.301	+1.02	[−0.22, 2.26]	0.109
	F1	+1.08	[−1.72, 3.89]	0.322	+1.05	[−0.21, 2.31]	0.064
Serous Fluid	ACC	+0.77	[−0.96, 2.50]	0.275	+0.96	[−1.48, 3.39]	0.426
	F1	+0.79	[−0.97, 2.56]	0.232	+0.95	[−1.49, 3.39]	0.557
Respiratory	ACC	+1.64	[−1.65, 4.93]	0.301	+0.77	[−0.64, 2.17]	0.297
	F1	+0.94	[−4.47, 6.36]	0.695	+0.56	[−1.37, 2.49]	0.695
Oral2026	ACC	+2.77	[−0.50, 6.04]	0.125	-0.86	[−4.46, 2.73]	0.875
	F1	-4.67	[−14.46, 5.11]	0.312	+1.02	[−7.27, 9.31]	1.00
Smart-CCS							
Cervical	ACC	+2.66	[0.80, 4.52]	0.023^{*}	+3.16	[0.12, 6.21]	0.037^{*}
	F1	+2.72	[0.78, 4.66]	0.027^{*}	+3.26	[0.15, 6.37]	0.037^{*}
Serous Fluid	ACC	+3.54	[1.70, 5.37]	0.004^{**}	+3.13	[1.37, 4.90]	0.004^{**}
	F1	+3.49	[1.66, 5.32]	0.002^{**}	+3.17	[1.40, 4.93]	0.004^{**}
Respiratory	ACC	+5.96	[4.13, 7.79]	0.002^{**}	+6.15	[3.16, 9.14]	0.006^{**}
	F1	+7.21	[4.23, 10.18]	0.002^{**}	+8.25	[4.01, 12.49]	0.010^{**}
Oral2026	ACC	-1.30	[−5.49, 2.90]	0.625	+2.80	[−1.13, 6.72]	0.188
	F1	+4.07	[−11.68, 19.83]	0.812	+3.43	[−3.15, 10.02]	0.188
UniCAS							
Cervical	ACC	+0.62	[−1.08, 2.31]	0.469	+0.51	[−0.70, 1.72]	0.570
	F1	+0.61	[−1.08, 2.31]	0.426	+0.55	[−0.69, 1.79]	0.359
Serous Fluid	ACC	+2.25	[−0.25, 4.74]	0.098	+1.91	[0.47, 3.36]	0.016^{*}
	F1	+2.26	[−0.23, 4.74]	0.105	+1.91	[0.46, 3.37]	0.020^{*}
Respiratory	ACC	+0.76	[−2.88, 4.40]	0.734	+1.15	[−2.10, 4.41]	0.232
	F1	+0.42	[−4.33, 5.18]	0.820	+1.78	[−2.28, 5.83]	0.322
Oral2026	ACC	+0.89	[−9.33, 11.10]	0.625	-0.22	[−3.97, 3.52]	1.00
	F1	+5.41	[−2.24, 13.05]	0.125	+2.75	[−4.85, 10.35]	0.438

A.7 Detailed Cross-Validation Results for Representative MIL Baselines

The compact main-text comparison reports method means to keep the three-backbone table readable.The following tables provide mean and standard deviation across cross-validation folds for methods for which fold-wise statistics are available in the supplied experimental results.

Table 14: Detailed comparison with existing MIL methods on the *Cervical* and *Serous Fluid* datasets. Results are reported as mean with standard deviation in parentheses. Best mean results are shown in **bold**, and second-best mean results are underlined.

Method	Cervical			Serous Fluid		
	AUC	ACC	F1	AUC	ACC	F1
DINOv3						
ABMIL	91.40 (2.11)	84.56 (3.45)	84.40 (3.44)	85.37 (2.57)	78.76 (2.23)	78.70 (2.27)
ACMIL	93.83 (1.96)	87.22 (3.33)	87.01 (3.43)	86.11 (3.49)	79.13 (3.21)	79.11 (3.21)
DTFD	92.29 (2.72)	86.00 (3.49)	85.66 (3.68)	85.13 (3.13)	78.06 (2.42)	77.99 (2.41)
ASMIL	92.57 (1.79)	87.02 (3.33)	86.89 (3.40)	84.31 (3.01)	77.40 (3.43)	77.32 (3.52)
DSMIL	90.00 (2.49)	83.84 (1.53)	83.59 (1.55)	83.06 (2.86)	76.52 (3.06)	76.46 (3.10)
TransMIL	91.32 (1.98)	83.44 (3.24)	83.21 (3.32)	85.03 (3.03)	77.40 (3.20)	77.35 (3.22)
MambaMIL	90.94 (2.73)	86.71 (3.61)	86.48 (3.76)	86.00 (2.37)	79.24 (1.95)	79.19 (1.97)
HMIL	91.93 (2.52)	84.15 (4.02)	83.84 (4.30)	84.78 (2.82)	77.14 (2.64)	77.03 (2.61)
PTCMIL	91.86 (2.11)	84.66 (2.99)	84.44 (3.02)	85.82 (2.92)	78.50 (2.80)	78.44 (2.85)
DI-ACMIL (Ours)	**93.86** (1.97)	**88.24** (3.12)	**88.06** (3.23)	**86.79** (3.50)	**80.09** (3.56)	**80.06** (3.56)
Smart-CCS						
ABMIL	94.38 (1.58)	87.42 (2.41)	87.28 (2.53)	84.22 (2.22)	76.52 (2.74)	76.49 (2.74)
ACMIL	94.31 (1.90)	86.82 (3.36)	86.61 (3.48)	85.04 (2.82)	77.51 (3.20)	77.43 (3.25)
DTFD	92.37 (1.69)	86.71 (3.03)	86.55 (2.99)	83.91 (3.03)	76.30 (2.77)	76.28 (2.77)
ASMIL	91.20 (2.38)	86.61 (2.25)	86.48 (2.34)	78.77 (3.99)	74.35 (2.37)	74.32 (2.39)
DSMIL	93.90 (2.30)	87.22 (3.32)	87.10 (3.35)	81.88 (2.71)	73.76 (2.83)	73.68 (2.90)
TransMIL	93.30 (1.32)	87.12 (3.40)	86.95 (3.57)	83.17 (3.25)	75.45 (2.73)	75.36 (2.75)
MambaMIL	92.92 (2.76)	87.74 (3.52)	87.65 (3.58)	83.00 (2.13)	75.52 (2.43)	75.46 (2.46)
HMIL	93.65 (0.80)	87.12 (2.16)	87.04 (2.22)	83.39 (2.59)	75.52 (3.14)	75.44 (3.17)
PTCMIL	94.61 (1.18)	87.02 (4.12)	86.82 (4.38)	84.87 (3.20)	76.59 (3.73)	76.46 (3.86)
DI-ACMIL (Ours)	**95.89** (1.21)	**89.98** (3.36)	**89.87** (3.35)	**88.33** (2.09)	**80.64** (2.62)	**80.60** (2.63)
UniCAS						
ABMIL	93.68 (2.43)	88.66 (2.36)	88.54 (2.39)	84.04 (2.21)	76.85 (2.04)	76.81 (2.01)
ACMIL	95.42 (0.74)	89.06 (1.73)	88.84 (1.85)	84.45 (2.40)	76.52 (2.27)	76.48 (2.26)
DTFD	92.66 (2.59)	86.30 (3.12)	86.09 (3.19)	82.76 (1.51)	74.97 (1.44)	74.93 (1.44)
ASMIL	90.58 (2.98)	87.43 (1.72)	87.30 (1.80)	80.63 (2.01)	74.31 (2.00)	74.28 (2.01)
DSMIL	92.17 (2.26)	86.20 (3.39)	85.98 (3.49)	80.31 (3.48)	73.17 (2.54)	73.08 (2.48)
TransMIL	92.06 (2.41)	84.97 (2.44)	84.78 (2.43)	80.18 (2.60)	73.17 (2.58)	73.12 (2.61)
MambaMIL	93.34 (1.93)	86.51 (2.79)	86.31 (2.82)	83.12 (2.32)	76.22 (2.00)	76.15 (2.03)
HMIL	93.55 (2.65)	87.02 (3.06)	86.91 (3.06)	82.04 (3.27)	75.08 (3.05)	74.95 (3.13)
PTCMIL	94.68 (1.64)	87.73 (3.49)	87.54 (3.70)	83.68 (2.53)	74.05 (2.94)	73.83 (3.04)
DI-ACMIL (Ours)	**95.79** (1.62)	**89.57** (2.14)	**89.39** (2.23)	**86.06** (2.41)	**78.43** (2.04)	**78.39** (2.01)

Table 15: Detailed comparison with existing MIL methods on the *Respiratory* and *Oral2026* datasets. Results are reported as mean with standard deviation in parentheses. Best mean results are shown in **bold**, and second-best mean results are underlined.

Method	Respiratory			Oral2026		
	AUC	ACC	F1	AUC	ACC	F1
DINOv3						
ABMIL	79.38 (4.18)	74.73 (4.14)	70.64 (5.31)	88.06 (5.07)	79.84 (2.85)	70.93 (5.04)
ACMIL	81.05 (4.66)	75.98 (3.15)	72.11 (3.87)	88.37 (5.15)	80.77 (2.89)	71.45 (5.90)
DTFD	75.77 (6.15)	71.57 (5.40)	66.90 (5.63)	78.98 (14.35)	81.50 (2.93)	58.29 (14.02)
ASMIL	73.11 (8.67)	74.54 (4.51)	69.95 (6.47)	85.02 (6.08)	80.65 (3.23)	58.73 (12.90)
DSMIL	74.42 (5.23)	72.43 (3.21)	68.16 (3.98)	85.60 (5.74)	82.67 (3.92)	**72.57** (4.25)
TransMIL	75.79 (6.36)	71.47 (4.57)	66.50 (6.57)	79.52 (8.52)	72.53 (7.49)	63.87 (6.74)
MambaMIL	80.38 (2.54)	74.06 (2.46)	70.34 (2.57)	83.22 (8.74)	**82.90** (2.91)	63.85 (11.98)
HMIL	79.19 (4.08)	74.16 (4.13)	70.28 (4.51)	82.16 (4.77)	81.51 (5.82)	65.58 (12.18)
PTCMIL	79.64 (5.12)	74.16 (3.73)	68.46 (5.46)	86.14 (5.42)	80.60 (4.93)	64.33 (11.68)
DI-ACMIL (Ours)	**82.01** (4.92)	**76.75** (4.19)	**72.67** (5.70)	**89.36** (4.53)	79.91 (4.09)	72.47 (5.20)
Smart-CCS						
ABMIL	79.36 (4.13)	74.16 (3.11)	70.04 (4.38)	83.91 (5.38)	83.18 (2.74)	67.18 (13.01)
ACMIL	81.21 (3.45)	76.17 (3.19)	71.83 (4.47)	88.20 (5.44)	81.73 (6.14)	72.26 (4.25)
DTFD	76.38 (5.15)	71.37 (4.93)	67.48 (5.31)	82.19 (10.55)	83.16 (2.72)	51.88 (14.73)
ASMIL	72.91 (5.59)	73.39 (3.62)	68.93 (5.32)	78.26 (6.09)	81.79 (4.41)	65.19 (11.63)
DSMIL	75.55 (5.27)	70.60 (5.38)	65.80 (6.25)	84.65 (5.31)	80.92 (8.60)	70.49 (6.48)
TransMIL	75.85 (5.09)	70.99 (4.85)	66.04 (6.32)	86.54 (7.64)	76.92 (2.92)	68.15 (5.16)
MambaMIL	79.16 (5.30)	74.07 (4.37)	70.11 (4.92)	84.32 (6.30)	83.61 (5.85)	67.57 (9.90)
HMIL	78.06 (3.71)	73.39 (4.08)	69.91 (4.42)	84.71 (2.73)	84.05 (4.59)	73.48 (6.74)
PTCMIL	79.13 (4.41)	73.39 (3.21)	69.00 (4.73)	87.79 (6.26)	84.08 (4.48)	73.75 (6.52)
DI-ACMIL (Ours)	**88.85** (2.07)	**82.32** (3.10)	**80.08** (3.48)	**89.13** (4.86)	**84.52** (5.25)	**75.69** (8.27)
UniCAS						
ABMIL	75.92 (4.99)	73.78 (3.39)	69.89 (4.28)	85.92 (4.96)	85.30 (7.65)	71.74 (8.80)
ACMIL	78.70 (2.84)	74.54 (2.70)	69.63 (3.91)	87.62 (3.40)	84.82 (1.27)	71.63 (6.63)
DTFD	70.90 (6.49)	68.30 (3.76)	63.07 (4.60)	85.06 (5.16)	**86.10** (4.42)	72.42 (9.78)
ASMIL	72.07 (4.77)	72.14 (4.14)	67.04 (5.71)	85.73 (3.21)	82.05 (2.85)	65.13 (10.62)
DSMIL	71.99 (6.03)	71.09 (5.57)	64.88 (8.34)	82.39 (2.39)	83.39 (5.42)	69.75 (5.38)
TransMIL	70.33 (6.07)	68.11 (6.71)	60.96 (8.32)	86.52 (3.37)	84.25 (2.65)	64.57 (12.20)
MambaMIL	76.02 (4.20)	71.18 (3.58)	65.98 (4.97)	83.22 (5.97)	82.20 (7.97)	70.03 (10.74)
HMIL	74.07 (5.96)	71.75 (4.60)	68.71 (4.98)	85.21 (2.45)	82.80 (4.95)	68.86 (1.68)
PTCMIL	76.25 (4.46)	72.14 (3.91)	65.70 (5.77)	86.31 (3.42)	83.71 (3.02)	70.24 (5.93)
DI-ACMIL (Ours)	**80.73** (4.05)	**75.70** (4.04)	**71.41** (4.46)	**88.76** (3.34)	84.60 (2.40)	**74.38** (5.66)

A.8 Cost Analysis by Storage

To complement the relative resource comparison in Figure 2, we report the absolute feature-storage requirements measured on the Oral2026 dataset, comprising 363 WSIs. All methods use 1024-dimensional Smart-CCS embeddings stored in float16 format. We distinguish between the storage required by the feature instances actually consumed by the MIL aggregator and the physical size of the cached feature files.

Storage of MIL input features. Table 16 reports the total number of MIL instances and their corresponding feature-storage requirements across different tile sizes. Conventional tile-based MIL retains one global [CLS] embedding per source tile, whereas DI-MIL constructs $k = 4$ decomposed instances from each original 2048-pixel tile. Each stored instance occupies $1024 \times 2 = 2048$ bytes.

Table 16: Absolute MIL feature-storage requirements on Oral2026 using Smart-CCS and ACMIL. Feature storage accounts for the float16 embeddings actually consumed by the MIL aggregator, excluding file-format overhead and auxiliary metadata. DI-MIL uses the original 2048-pixel tiles with $k = 4$. Performance is reported as mean AUC and macro-F1 (%).

Method	Instances	Storage (GiB)	AUC	Macro-F1
Tile 2048	0.94M	1.79	88.20	72.26
Tile 1200	2.67M	5.09	88.95	73.89
DI-MIL ($k = 4$)	3.75M	7.14	89.13	75.69
Tile 600	10.40M	19.83	88.73	74.40
Tile 300	40.52M	77.30	86.20	74.35

DI-MIL requires 7.14 GiB of MIL input features, compared with 1.79 GiB for the original Tile 2048 baseline. This fourfold increase reflects the expansion of the MIL bag from N to $4N$, rather than additional image extraction or encoder forward passes. Compared with Tile 1200, DI-MIL requires approximately 2.05 GiB of additional feature storage while improving macro-F1 by 1.80 percentage points. In contrast, direct fine-grid tiling at 600 and 300 pixels requires 19.83 and 77.30 GiB, respectively, while yielding lower AUC and macro-F1 than DI-MIL.

Physical storage of cached features. The actual cached feature files contain additional information beyond the embeddings consumed by MIL. Specifically, our feature-extraction pipeline stores one [CLS] token and four cluster embeddings per tile, together with auxiliary cluster-assignment maps. The same 2048-pixel feature cache is shared by conventional tile-based MIL and DI-MIL; the latter selects the four cluster embeddings instead of the [CLS] embedding and therefore does not require a separate copy of the source features.

Across all 363 WSIs, the 2048-pixel feature cache occupies 9.17 GiB on disk, including 8.93 GiB of stored token embeddings and 0.22 GiB of cluster-assignment maps. The corresponding feature caches for 1200-, 600-, and 300-pixel tiling occupy 26.13, 101.80, and 396.74 GiB, respectively. These physical file sizes include four cluster embeddings per tile that are not used by the conventional single-instance baselines. Therefore, the MIL input-feature sizes in Table 16 provide a more representative comparison of the feature storage required by each instance-construction strategy, whereas the physical file sizes reflect our shared experimental caching format.

Why not directly use all dense tokens as MIL instances? A natural alternative to DI-MIL is to treat every dense spatial token produced by the frozen encoder as an independent MIL instance. However, this would substantially increase the storage and computational requirements of slide-level aggregation. As reported in Table 7, Smart-CCS produces 256 dense spatial tokens per encoded tile. Directly exposing all tokens to the MIL aggregator would therefore expand the slide-level bag from N to $256N$ instances, compared with only $4N$ instances in DI-MIL.

For Oral2026, retaining all dense tokens as float16, 1024-dimensional MIL instances would require approximately 456.96 GiB of feature storage, compared with 7.14 GiB for DI-MIL. This represents a 64-fold increase in both the number of MIL instances and their feature-storage requirements, despite using exactly the same source tiles and encoder forward passes. The actual downstream

memory footprint would be further affected by intermediate activations, attention computations, and model-specific training requirements.

Importantly, DI-MIL performs token clustering independently within each source tile and caches only the resulting four instance embeddings. In contrast, directly using all dense tokens would propagate the complete token sequence into slide-level learning, substantially increasing feature storage, data-loading overhead, and the computational burden of downstream aggregation. This cost becomes particularly consequential for large-scale cytopathology cohorts containing thousands of WSIs.

These results motivate decomposed-instance construction as a practical intermediate strategy: DI-MIL preserves multiple independently weighted sub-tile representations while reducing the number of instances exposed to the MIL aggregator by a factor of 64 relative to direct dense-token aggregation.

A.9 Cost Analysis by Computational Time

While the main text evaluates cost efficiency from a storage perspective, here we analyze the empirical runtime requirements (GPU-hours) of DI-MIL compared against conventional fine-grid tiling strategies on the Respiratory dataset ($n = 1041$ slides).

As shown in Figure 6b, physically reducing tile size introduces severe computational bottlenecks during feature extraction. Compared to the baseline Tile 1200 ($1\times$), cropping finer tiles escalates extraction runtime drastically: Tile 600 requires $3.91\times$ GPU-hours, while Tile 300 inflates feature extraction time by $20.15\times$ (~ 58.4 GPU-hours). In contrast, DI-MIL ($k = 4$) performs lightweight token clustering in feature space post-encoding, incurring a negligible total runtime penalty of only $1.04\times$ the baseline extraction cost.

Furthermore, DI-MIL achieves a superior trade-off between diagnostic performance and computational efficiency (Figure 6a). While fine-grid tiling yields modest AUC gains at exponential computational expense (85.5% AUC at $20.15\times$ compute for Tile 300), DI-MIL ($k = 4$) achieves a significantly higher AUC of 88.8% ($+7.6\%$ over Tile 1200). It matches the $4\times$ bag length of Tile 600 while maintaining virtually the same extraction cost as Tile 1200. This confirms that DI-MIL provides Pareto-dominant efficiency, eliminating the heavy GPU forward-pass overhead required by traditional sub-tiling.

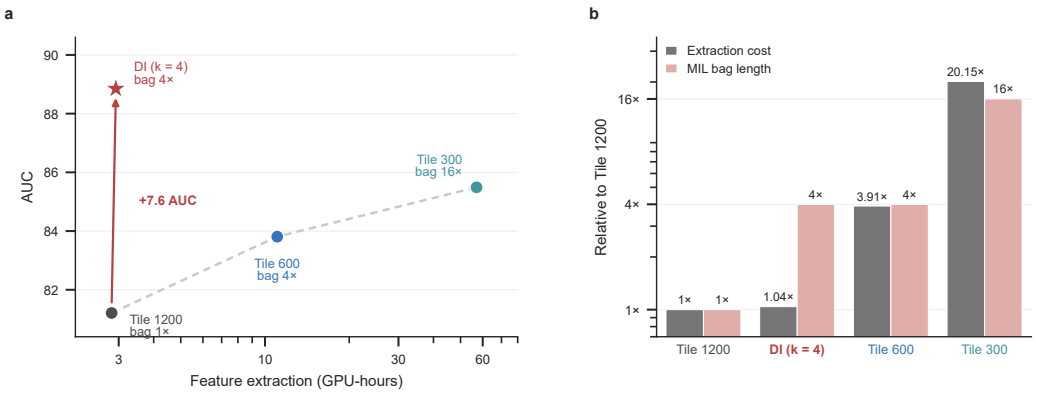


Figure 6: **Feature-extraction cost, MIL bag length, and AUC on Respiratory. (a)** AUC versus measured CCS feature-extraction GPU-hours for conventional tiles (dashed) and DI ($k = 4$). Marker annotations give the median MIL bag relative to Tile 1200. **(b)** The same methods, split into extraction cost and MIL bag length, both relative to Tile 1200. DI matches the 4× bag of Tile 600 while remaining at Tile-1200 extraction cost. $n = 1041$ slides; MIL times are isolated ACMIL steps on a RTX 3090 GPU and are not taken from the contended training wall-clock.