CELL2CASE: A UNIFIED MULTI-TASK MODEL FROM CELL-LEVEL PERCEPTION TO PATIENT-LEVEL DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Patient-level hematopathology diagnosis requires integrating diverse cellular observations distributed across multiple patient-associated microscopic views. Existing unified models integrate multiple cell- and image-level tasks, but their unification remains largely task-wise and does not explicitly organize local predictions into patient-level diagnostic evidence. To address this problem, we propose Cell2Case, a unified framework for cell-to-case prediction. Cell2Case enriches detector-anchored queries with multi-scale context through gated residual refinement, then binds class, location, morphology, and contextual features into query-aligned cellular records. These records support perception and language tasks and are hierarchically aggregated into view- and patient-level representations. The resulting patient evidence preserves salient cellular events, expected cell-type counts, morphology statistics, and relationships across views rather than collapsing all observations into a generic pooled representation. We evaluate Cell2Case on seven tasks using 46 dataset configurations from 36 public hematopathology collections and three patient cohorts. Under a shared evaluation protocol, Cell2Case achieves the highest run-averaged score among the evaluated baselines on every supported task. On seven-class diagnosis using a patient-disjoint test set pooled from three cohorts, it improves balanced accuracy by 17.84 percentage points over the strongest evaluated patient-level baseline. Code is available at `https://anonymous.4open.science/r/Cell2Case-F802`.

1 INTRODUCTION

Digital hematopathology has advanced at the level of individual cells. Deep models can recognize malignant leukocytes, differentiate fine-grained bone-marrow cell types, and identify disease-associated cytomorphological patterns (Matek et al., 2019; 2021; Sidhom et al., 2021; de Almeida et al., 2023). Recent systems demonstrate that cell-level models can generalize across institutions and acquisition settings (Sun et al., 2025). Yet patient-level diagnosis is not obtained by classifying cells independently. Diagnostic evidence can be *sparse*, when only a few abnormal cells are informative; *compositional*, when disease is reflected in altered cell-type proportions; or *distributed*, when morphological variation becomes apparent only across multiple microscopic views. These properties create a scale mismatch between local prediction and case-level reasoning: evidence must remain cell-specific, preserve population-level composition and morphology, and be integrated across views without allowing common observations to overwhelm sparse but informative events.

Recent work has increasingly unified traditionally separate tasks within digital hematopathology. Morphology-aware supervision exposes interpretable nuclear, cytoplasmic, and granular attributes (Tsutsui et al., 2023), while hematology-specific self-supervised pretraining improves the transferability of cellular representations (Koch et al., 2024). More recently, Uni-Hema unifies detection, classification, segmentation, morphology prediction, and language reasoning within a single multi-task architecture (Rehman et al., 2026). These advances substantially broaden what a single model can infer from a microscopic image, but their unification remains primarily task-wise rather than case-wise. Local outputs are optimized for their respective tasks but are not explicitly organized into structured cellular evidence that can be reused and accumulated across patient-associated views. The central challenge is therefore to extend unification across biological scales, from cell-level

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.

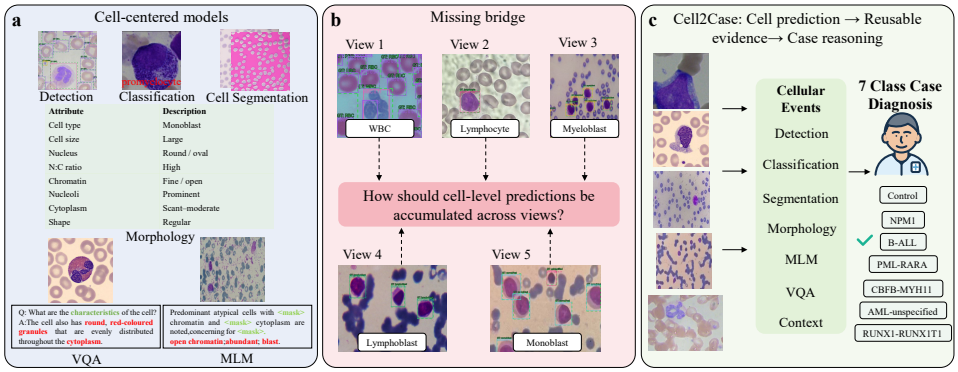


Figure 1: **The cell-to-case interface.** Cell2Case preserves query-aligned local predictions as structured cellular evidence and hierarchically aggregates them across patient-associated views.

prediction to patient-level reasoning, while preserving the identity, localization, morphology, and contextual semantics learned locally.

Patient-level aggregation addresses a complementary part of this problem. Attention-based multiple-instance learning and recent hematology-specific systems show that case labels can identify diagnostically informative cells and that aggregating large cellular collections is valuable (Ilse et al., 2018; Manescu et al., 2023; Singi et al., 2026). However, these approaches primarily optimize a case-level endpoint from instance embeddings rather than preserving heterogeneous predictions from a unified cellular model as reusable evidence. The unresolved problem is therefore not simply instance aggregation, but a prediction-to-evidence interface that preserves task-aligned cellular information and organizes it across views.

To address this gap, we propose **Cell2Case**, a unified framework that establishes a structured cell-to-case pathway. At the cellular level, an anchor-preserving residual refiner enriches selected detector queries with multi-scale context via gated residual updates anchored to their initial predictions. Cell2Case then binds refined class, box, and morphology attributes with expert-transformed contextual embeddings by query identity, forming class–box–morphology–event (CBME) records. At the patient level, a hierarchical patient evidence memory (HPEM) converts these records into view summaries using sparse-evidence attention, expected cell-type counts, and morphology statistics before modeling their interactions across views. Rather than collapsing cellular observations into a generic pooled representation, this pathway preserves their structured semantics as evidence is accumulated from cells to views and ultimately to the patient.

We evaluate Cell2Case across seven perception, language, and diagnosis tasks using 46 dataset configurations from 36 public hematopathology collections and three patient cohorts comprising 396 cases. Under the shared evaluation protocol, Cell2Case achieves the highest run-averaged score among the evaluated baselines on every supported task. Multi-task training improves all six non-detection tasks over their independently trained counterparts while maintaining detection performance. On patient-disjoint seven-class diagnosis pooled across three cohorts, Cell2Case reaches 75.01% balanced accuracy, improving by 17.84 percentage points over the strongest evaluated patient-level baseline. Ablation and view-scaling analyses support the roles of cellular evidence quality and cross-view evidence organization.

Our contributions are threefold:

- We formulate unified hematopathology as a cell-to-case learning problem and identify the structured evidence interface between local multi-task prediction and patient-level reasoning.
- We introduce an anchor-preserving, query-aligned cellular representation that retains the correspondence among identity, localization, morphology, and contextual evidence.

- We develop hierarchical evidence aggregation for sparse, compositional, and distributed diagnostic signals, and validate the resulting framework across seven tasks and three patient cohorts.

2 Related Work

Cell-level perception and representation. Computational hematology has traditionally decomposed microscopic analysis into classification, detection, segmentation, and morphology prediction. Public peripheral-blood and bone-marrow resources provide the corresponding cell labels, locations, masks, and expert attributes (Acevedo et al., 2020; Kouzehkanan et al., 2022; Matek et al., 2021; Tsutsui et al., 2023), while disease-oriented systems show that cellular appearance supports blast recognition, leukemia screening, and morphotype discovery (Matek et al., 2019; Manescu et al., 2023; de Almeida et al., 2023). These methods enable precise local perception but often specialize representations to a particular dataset or objective. Self-supervised learning improves transferability: DINOv2 provides general visual features, and DinoBloom adapts this paradigm to hematological cells (Oquab et al., 2024; Koch et al., 2024). Nevertheless, a transferable cell embedding still describes an individual cell or image; it does not specify how task-aligned predictions should become evidence for a patient case.

Unified cellular and multimodal models. Pathology vision–language models such as PLIP and CONCH align visual features with biomedical text (Huang et al., 2023; Lu et al., 2024). Within hematology, WBCAtt introduces expert morphology supervision, whereas Uni-Hema combines detection, classification, segmentation, morphology prediction, VQA, and MLM (Tsutsui et al., 2023; Rehman et al., 2026). This line of work replaces isolated task models with shared visual and language representations, enriching what can be inferred from a cell or image. Such unification, however, remains task-wise rather than case-wise: predictions from multiple patient-associated views are not explicitly preserved as a common evidence unit for diagnosis.

Patient-level aggregation. Multiple-instance learning addresses the complementary problem by constructing case representations from unordered visual instances (Ilse et al., 2018; Campanella et al., 2019; Lu et al., 2021). In hematology, related approaches support weakly supervised leukemia classification, morphotype-based analysis, and diagnosis from large cell collections (Manescu et al., 2023; de Almeida et al., 2023; Koch et al., 2024; Singi et al., 2026; Dasdelen et al., 2026). These studies establish that population-level cellular structure is diagnostically informative, but typically optimize a case-level endpoint over generic or precomputed cell embeddings. Consequently, case aggregation remains largely separated from the localization, morphology, dense-prediction, and language semantics learned by unified cellular models. Cell2Case targets this missing interface by converting query-aligned cellular predictions into structured records that can be reused across tasks and accumulated across patient-associated microscopic views.

3 Method

3.1 Overview and Visual Evidence

Given a microscopic view $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, Cell2Case supports detection, classification, segmentation, morphology prediction, visual question answering (VQA), and masked language modeling (MLM). Depending on the source dataset, a view may be a complete microscopic field or a cell-centered image. Patient diagnosis operates on a collection of patient-associated views $\mathcal{X}_p = \{\mathbf{x}_{p,f}\}_{f=1}^{F}$. We use f and i to index views and cell queries, respectively.

As illustrated in Figure 2, Cell2Case follows a cell-to-case hierarchy. It first extracts multi-scale visual features and refines selected detector queries. The refined attributes and contextual event embeddings are then bound into query-aligned cellular records. These records support cellular and language tasks and are hierarchically aggregated across views for seven-class patient diagnosis.

Multi-scale visual evidence. Cellular analysis requires both fine spatial detail and broader semantic context. A trainable convolutional stream extracts local feature maps $\mathbf{F}_4$, $\mathbf{F}_8$, and $\mathbf{F}_{16}$ at strides 4, 8, and 16. A frozen DINOv2 encoder (Oquab et al., 2024) provides semantic patch features $\mathbf{S}$ and a global token $\mathbf{z}_{\text{cls}}$. At each scale, resized semantic features $\mathbf{S}_s$ are selectively fused with local

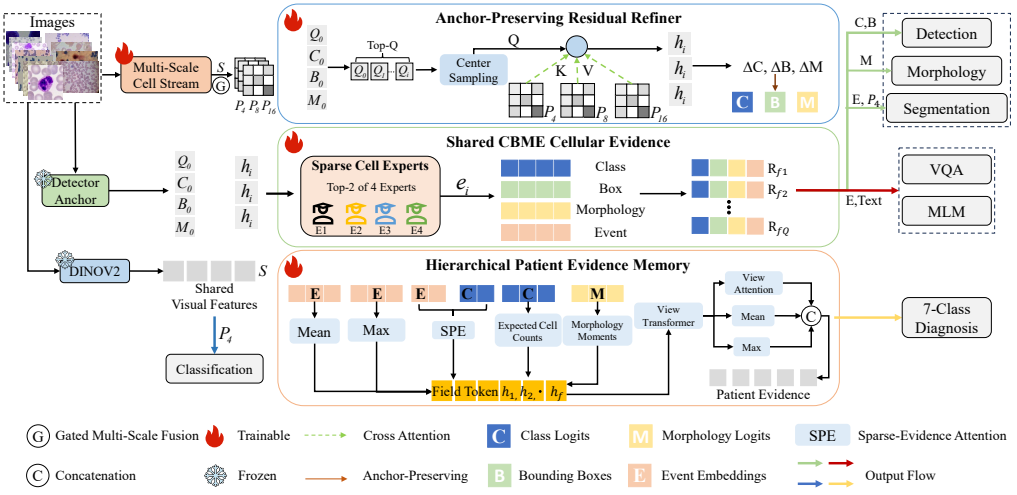


Figure 2: **Overview of Cell2Case.** Anchor-preserving refinement updates class, box, and morphology attributes, while sparse experts derive event embeddings; these query-aligned signals form CBME records. Task-specific branches consume the corresponding evidence, while HPEM aggregates cellular records across patient-associated views for diagnosis.

features before top-down aggregation:

$$\begin{aligned} \mathbf{U}_s &= \psi_s(\mathbf{F}_s + \sigma(g_s([\mathbf{F}_s; \mathbf{S}_s])) \odot \mathbf{S}_s)\,, \\ \mathbf{P}_{16} &= \mathbf{U}_{16}, \qquad \mathbf{P}_8 = \mathbf{U}_8 + \mathrm{Up}(\mathbf{P}_{16}), \\ \mathbf{P}_4 &= \mathbf{U}_4 + \mathrm{Up}(\mathbf{P}_8). \end{aligned} \tag{1}$$

Here, $[\cdot;\cdot]$ denotes concatenation, $\odot$ denotes element-wise multiplication, and ψ_s is a convolutional smoothing block. The resulting pyramid $\mathbf{P} = \{\mathbf{P}_4, \mathbf{P}_8, \mathbf{P}_{16}\}$ preserves high-resolution cellular boundaries while propagating semantic context across scales. Task-specific residual adapters operate on $\mathbf{P}_4$ and are initialized to produce zero residuals.

3.2 Anchor-Preserving Residual Refinement

The released Uni-Hema model (Rehman et al., 2026) is used as a frozen cellular anchor, providing stable query-aligned class, box, and morphology hypotheses. Cell2Case refines its selected queries while explicitly preserving the anchor predictions at initialization. The frozen anchor produces queries $\mathbf{Q}_0$, class logits $\mathbf{C}_0$, normalized boxes $\mathbf{B}_0$, and morphology logits $\mathbf{M}_0$. We retain the top Q queries according to maximum sigmoid class confidence and apply the same indices to all four outputs. This selection preserves the correspondence between each query and its class, location, and morphology predictions.

For each selected query, we bilinearly sample $\mathbf{f}_i^{\mathrm{center}}$ from $\mathbf{P}_4$ at the center of its anchor box. The projected anchor query and sampled feature form a localized query that attends to tokens from the complete feature pyramid:

$$\begin{aligned} \mathbf{q}_i &= \phi_q(\mathbf{Q}_{0,i}) + \phi_l(\mathbf{f}_i^{\mathrm{center}}), \\ \mathbf{K} = \mathbf{V} &= [\mathrm{Flat}(\mathbf{P}_4); \mathrm{Flat}(\mathbf{P}_8); \mathrm{Flat}(\mathbf{P}_{16})], \\ \mathbf{h}_i &= \mathrm{LN}(\mathbf{q}_i + \mathrm{MHA}(\mathbf{q}_i, \mathbf{K}, \mathbf{V}))\,. \end{aligned} \tag{2}$$

The learned projections ϕ_q and ϕ_l map detector queries and sampled features into the shared d-dimensional space. Center sampling retains local cellular appearance, whereas cross-attention incorporates surrounding cells and tissue context. The resulting $\mathbf{h}_i$ remains anchored to a detected cell but contains information from the entire view.

Three prediction heads map $\mathbf{h}_i$ to residual class, box, and morphology updates. A shared learnable scalar controls their contribution:

$$\begin{aligned} \mathbf{c}_i &= \mathbf{c}_{0,i} + \alpha \Delta \mathbf{c}_i, \\ \mathbf{b}_i &= \sigma\big(\sigma^{-1}(\mathbf{b}_{0,i}) + 0.1\alpha \Delta \mathbf{b}_i\big), \\ \mathbf{m}_i &= \mathbf{m}_{0,i} + \alpha \Delta \mathbf{m}_i, \qquad \alpha = \tanh(g), \quad g_{\text{init}} = 0. \end{aligned} \tag{3}$$

Because $\alpha = 0$ initially, the refined predictions exactly equal the selected anchor outputs before adaptation. Training subsequently learns gated residual updates without requiring the model to reconstruct the anchor predictions from scratch. Box updates are applied in inverse-sigmoid space to preserve valid normalized coordinates. The residual heads use ordinary initialization so that gradients reach the scalar gate from the first optimization step.

3.3 Shared Cellular Evidence and Multi-Task Prediction

The refined attributes alone do not provide a common representation for dense prediction, language reasoning, and patient aggregation. Cell2Case therefore derives a contextual event embedding from every refined query. A softmax router assigns each query to four sparse experts, of which the two highest-weighted experts contribute:

$$\begin{aligned} \boldsymbol{\pi}_{f,i} &= \mathrm{Softmax}(\mathbf{W}_r \mathbf{h}_{f,i} + \mathbf{b}_r), \\ \mathbf{e}_{f,i} &= \mathbf{h}_{f,i} + \sum_{r \in \mathrm{Top2}(\boldsymbol{\pi}_{f,i})} \pi_{f,i,r} \mathcal{E}_r(\mathbf{h}_{f,i}), \\ \mathbf{R}_{f,i} &= \{\mathbf{c}_{f,i}, \mathbf{b}_{f,i}, \mathbf{m}_{f,i}, \mathbf{e}_{f,i}\}, \qquad \mathcal{R}_f = \{\mathbf{R}_{f,i}\}_{i=1}^{Q}. \end{aligned} \tag{4}$$

Here, $\mathcal{E}_r$ denotes expert r. Each class–box–morphology–event (CBME) record $\mathbf{R}_{f,i}$ binds cell identity, location, morphology, and contextual features through a common query index. The record is not an additional prediction head. Instead, it provides query-aligned cellular evidence for downstream prediction and patient aggregation. Sparse routing allows event features to specialize without separating them into independent task representations.

Cellular and image prediction. Detection consumes $(\mathbf{C}, \mathbf{B})$, while morphology prediction consumes $\mathbf{M}$, which contains two logits for each of six binary attributes. Image classification uses the concatenation of $\mathbf{z}_{\text{cls}}$ and average- and max-pooled $\mathbf{P}_4$. Classification therefore shares the visual pathway but does not directly consume CBME. Segmentation combines high-resolution spatial features with cell-aligned event context. A six-way segmentation-ontology ID provided with each segmentation record is embedded into channel-wise scale and shift parameters that condition projected $\mathbf{P}_4$ features. Inner products between conditioned pixels and projected event embeddings produce query-specific masks. Their log-sum-exp aggregation is combined with a semantic foreground head, while a parallel head predicts cell boundaries. This design allows segmentation to retain pixel-level detail while using event embeddings to distinguish cellular structures.

Event-grounded language prediction. BioClinicalBERT (Alsentzer et al., 2019) encodes questions or masked clinical descriptions, and a learned projection maps the event embeddings into its hidden space. For VQA, pooled text and event features support question-type prediction, structured answers, and image–text matching. A Transformer decoder generates open-ended answers by attending jointly to question tokens and cellular events. For MLM, contextualized text tokens attend to the same event embeddings before vocabulary prediction. The output layer reuses the pretrained language model's prediction transform, word embeddings, and vocabulary bias. Thus, VQA and MLM share cellular evidence while retaining task-specific output mechanisms. Detailed inference rules are provided in Appendix A.3.

3.4 Hierarchical Patient Evidence Memory

We define image-derived patient-level diagnostic evidence as a case representation that preserves salient cellular events, expected cell-type counts, morphology statistics, and relationships among patient-associated views. Global pooling can suppress rare abnormalities and obscure population-level shifts. HPEM instead constructs this evidence through event-to-view and view-to-patient aggregation. HPEM directly aggregates event embeddings, class logits, and morphology logits, while spatial context enters indirectly through localization and center sampling.

Event-to-view aggregation. Mean and max pooling capture typical and extreme event features. Sparse-evidence attention combines a learned event score with detector confidence, allowing salient, high-confidence events to contribute more strongly. Expected cell-type counts summarize composition by summing sigmoid class probabilities, while morphology is represented by the mean and variance of positive-class probabilities:

$$\begin{aligned} a_{f,i} &= \mathrm{Softmax}_i\left[\mathbf{w}_a^\top \mathbf{e}_{f,i} + \log\left(\max\{\epsilon, \max_k \sigma(c_{f,i,k})\}\right)\right], \\ \mathbf{u}_f &= [\mathrm{Mean}_i(\mathbf{e}_{f,i}); \mathrm{Max}_i(\mathbf{e}_{f,i}); \sum_i a_{f,i}\mathbf{e}_{f,i}; \sum_i \sigma(\mathbf{c}_{f,i}); \boldsymbol{\mu}_f^M; \boldsymbol{\nu}_f^M], \\ \mathbf{h}_f &= \phi_F(\mathbf{u}_f). \end{aligned} \tag{5}$$

Here, ϕ_F is an affine projection followed by GELU and layer normalization, $\epsilon = 10^{-5}$, and $\boldsymbol{\mu}_f^M, \boldsymbol{\nu}_f^M$ denote morphology means and variances. The resulting view token $\mathbf{h}_f \in \mathbb{R}^d$ therefore represents event content, event salience, cellular composition, and morphological variation. Sparse-evidence attention requires no explicit event-salience annotations, and expected counts are probabilistic summaries rather than laboratory-calibrated blood counts.

View-to-patient aggregation. A View Transformer contextualizes all F view tokens so that evidence from one view can modify the interpretation of another. Mean pooling, max pooling, and learned view attention then capture complementary global summaries:

$$\begin{aligned} \widetilde{\mathbf{H}}_p &= \mathrm{Transformer}([\mathbf{h}_{p,1}, \ldots, \mathbf{h}_{p,F}]), \\ \beta_{p,f} &= \mathrm{Softmax}_f(\mathbf{w}_{\mathrm{att}}^\top \widetilde{\mathbf{h}}_{p,f}), \\ \mathbf{z}_p &= [\mathrm{Mean}_f(\widetilde{\mathbf{h}}_{p,f}); \mathrm{Max}_f(\widetilde{\mathbf{h}}_{p,f}); \sum_f \beta_{p,f}\widetilde{\mathbf{h}}_{p,f}]. \end{aligned} \tag{6}$$

The patient representation $\mathbf{z}_p \in \mathbb{R}^{3d}$ feeds a seven-class diagnostic head and auxiliary heads for normal versus abnormal status and myeloid versus lymphoid lineage. A source classifier connected through gradient reversal discourages the representation from encoding cohort-specific acquisition cues.

3.5 Learning Objectives and Training

Cellular and multimodal supervision. Detection uses Hungarian matching based on classification, box ℓ_1, and generalized-IoU costs. Morphology loss is computed only for annotated attributes. To prevent excessive drift after initialization, the detection objective includes a penalty relative to the selected anchor predictions:

$$\begin{aligned} \mathcal{L}_{\mathrm{det}} &= 2\mathcal{L}_{\mathrm{focal}} + 5\mathcal{L}_{\mathrm{box}} + 2\mathcal{L}_{\mathrm{GIoU}} + \mathcal{L}_{\mathrm{morph}} + 0.25\mathcal{L}_{\mathrm{anchor}}, \\ \mathcal{L}_{\mathrm{anchor}} &= \mathrm{Mean}[(\mathbf{C} - \mathbf{C}_0)^2] + 5\,\mathrm{Mean}[|\mathbf{B} - \mathbf{B}_0|]. \end{aligned} \tag{7}$$

Zero gating preserves the anchor exactly at initialization, whereas $\mathcal{L}_{\mathrm{anchor}}$ constrains subsequent adaptation. Classification uses Balanced-Softmax focal loss, and segmentation combines binary cross-entropy, Dice, and boundary losses. VQA jointly optimizes structured answers, generated answers, question type, and image–text matching. MLM combines masked-token prediction with text-only and shuffled-event contrast conditions.

Patient supervision. Patient training combines seven-class diagnosis with coarse diagnostic supervision, hierarchy consistency, and source-adversarial learning:

$$\mathcal{L}_{\mathrm{case}} = \mathcal{L}_{7\mathrm{cls}} + 0.5\mathcal{L}_{\mathrm{normal}} + 0.35\mathcal{L}_{\mathrm{lineage}} + 0.2\mathcal{L}_{\mathrm{hier}} + 0.03\mathcal{L}_{\mathrm{source}}. \tag{8}$$

Coarse labels are derived from diagnosis, excluding Control from lineage supervision; the hierarchy loss aligns abnormality probabilities, while $\mathcal{L}_{\mathrm{source}}$ applies gradient reversal.

Training progressively optimizes residual refinement, dense visual prediction, and language and patient reasoning through a staged schedule. The anchor remains frozen throughout, while staged optimization limits interference with previously learned visual pathways. The complete module-wise freezing schedule and learning rates are reported in Appendix A.3, Table 6.

Table 1: Comparison on fixed, training-disjoint evaluation partitions (mean_{std}, %) over three runs. "–" denotes an unsupported task.

Method	Detection mAP_{50}	Classification Bal. Acc.	Segmentation Dice	Morphology CA-Macro-F1	VQA Answer Token-F1	MLM Token Acc.	Diagnosis Bal. Acc.
Task-specific models							
RT-DETR (Zhao et al., 2024)	$26.80_{0.42}$	–	–	–	–	–	–
ResNet-50 (He et al., 2016)	–	$62.79_{0.51}$	–	–	–	–	–
DinoBloom-S (Koch et al., 2024)	–	$73.18_{0.35}$	–	–	–	–	–
U-Net (Ronneberger et al., 2015)	–	–	$87.19_{0.28}$	–	–	–	–
TransUNet (Chen et al., 2021)	–	–	$87.33_{0.31}$	–	–	–	–
AttriDet (Rehman et al., 2024)	$17.68_{0.39}$	–	–	$47.70_{0.62}$	–	–	–
Cell2Case (single-task variants)	$66.12_{0.21}$	$73.77_{0.29}$	$90.66_{0.18}$	$90.37_{0.25}$	$62.94_{0.45}$	$75.96_{0.33}$	$70.55_{1.28}$
Medical vision–language models							
M^3AE (Chen et al., 2022)	–	–	–	–	$24.75_{0.58}$	$69.11_{0.41}$	–
BLIP-2 (Li et al., 2023)	–	–	–	–	$35.01_{0.63}$	–	–
Patient-level aggregation models							
ABMIL (Ilse et al., 2018)	–	–	–	–	–	–	$56.02_{1.85}$
TransMIL (Shao et al., 2021)	–	–	–	–	–	–	$51.50_{1.92}$
MILLIE (Manescu et al., 2023)	–	–	–	–	–	–	$52.32_{1.78}$
CAREMIL (Singi et al., 2026)	–	–	–	–	–	–	$57.17_{4.64}$
cAItomorph (Dasdelen et al., 2026)	–	–	–	–	–	–	$54.17_{1.71}$
Unified models							
Uni-Hema (Rehman et al., 2026)	$62.43_{0.31}$	$55.85_{0.46}$	$86.35_{0.37}$	$82.25_{0.50}$	$70.24_{0.28}$	$75.53_{0.35}$	–
Cell2Case (multi-task)	$\mathbf{66.19}_{0.15}$	$\mathbf{80.55}_{0.18}$	$\mathbf{92.94}_{0.12}$	$\mathbf{92.07}_{0.16}$	$\mathbf{72.68}_{0.22}$	$\mathbf{77.87}_{0.19}$	$\mathbf{75.01}_{3.39}$

4 Experiments

4.1 Experimental Setup

We report mAP_{50} for detection, balanced accuracy for classification and diagnosis, Dice for segmentation, CA-Macro-F1 for morphology, Answer Token-F1 for VQA, and masked-token accuracy for MLM. All metrics are percentages. The cellular and language benchmark contains 46 dataset configurations drawn from 36 public collections. Dataset composition is provided in Appendix A.1, while metric definitions and implementation details are provided in Appendix A.3.

Diagnosis is a harmonized seven-class pooled-cohort task over JHU APL/AML, MLL Helmholtz, and C-NMC 2019. Their label support is not fully crossed, so we report balanced accuracy and audit source dependence. The 396 patients are split 274/54/68 for training/validation/test; the 122 validation and test patients and all associated records are excluded from every training partition. Cell2Case inputs are resized to 336×336. Cell2Case uses $d = 192$, at most 300 queries, four top-2 experts, and up to 32 views. AdamW uses weight decay 0.05, cosine decay, at most 500 warm-up steps, stage learning rates $\{3, 5, 2, 1\} \times 10^{-5}$, and batches of two images or one patient bag.

4.2 Comparison Results

Table 1 compares Cell2Case with task-specific, vision–language, patient-level, and unified baselines under this shared evaluation protocol. Cell2Case achieves a mean detection mAP_{50} of 66.19, classification balanced accuracy of 80.55, segmentation Dice of 92.94, morphology CA-Macro-F1 of 92.07, VQA Answer Token-F1 of 72.68, and MLM masked-token accuracy of 77.87. It outperforms Uni-Hema on all six shared tasks, including a 3.76-point gain in detection. The larger gains in classification, segmentation, and morphology are consistent with complementary contributions from shared visual transfer and query-aligned cellular evidence: classification relies on the shared visual pathway, whereas segmentation and morphology directly exploit cell-aligned representations.

For patient-level diagnosis, Cell2Case achieves 75.01 mean balanced accuracy on the held-out 68-patient test set, compared with 57.17 for CAREMIL. The paired class-stratified bootstrap estimates a 17.84-point improvement with a 95% confidence interval of $[8.64, 27.19]$. This result measures patient-disjoint pooled-cohort classification rather than unseen-cohort generalization. Multi-task training improves all six non-detection tasks over the corresponding single-task variants, with gains of 9.74 points in VQA and 6.78 points in classification, while detection remains stable at $66.19_{0.15}$ compared with $66.12_{0.21}$. This pattern is consistent with positive transfer from shared supervision while maintaining localization performance. Source-only and uncertainty analyses are provided in Appendix A.5.

Table 2: Component ablations of Cell2Case (mean$_{\text{std}}$, %) over three runs. The residual-update ablation retains contextual event embeddings but uses anchor attributes (C_0, B_0, M_0). Sparse experts and HPEM are replaced by a shared feed-forward network and global mean pooling, respectively. Dashes denote unaffected task paths.

Variant	Detection mAP$_{50}$	Classification Bal. Acc.	Segmentation Dice	Morphology CA-Macro-F1	VQA Answer Token-F1	MLM Token Acc.	Diagnosis Bal. Acc.
w/o residual updates	$62.43_{0.32}$	–	$89.65_{0.25}$	$89.81_{0.30}$	$72.04_{0.21}$	$77.35_{0.23}$	$71.82_{1.35}$
w/o zero gating	$64.88_{0.29}$	–	–	$90.94_{0.26}$	–	–	$71.95_{1.28}$
w/o sparse experts	–	–	$90.88_{0.27}$	–	$70.12_{0.38}$	$75.31_{0.32}$	$69.84_{1.42}$
w/o HPEM	–	–	–	–	–	–	$59.40_{1.62}$
Full Cell2Case	$\mathbf{66.19}_{0.15}$	$\mathbf{80.55}_{0.18}$	$\mathbf{92.94}_{0.12}$	$\mathbf{92.07}_{0.16}$	$\mathbf{72.68}_{0.22}$	$\mathbf{77.87}_{0.19}$	$\mathbf{75.01}_{3.39}$

4.3 Ablation Studies

Component contributions. Table 2 reveals distinct roles for local attribute refinement and event specialization. Replacing the sparse experts with a shared feed-forward network reduces segmentation by 2.06 points, VQA and MLM by 2.56 points, and diagnosis by 5.17 points. The consistent degradation across pixel-, language-, and patient-level tasks suggests that expert routing is most valuable when the same cellular events must support heterogeneous downstream reasoning. Rather than improving a single prediction head, it appears to increase the adaptability of the shared event representation while preserving a common query-aligned space.

Disabling residual updates produces a different pattern. Detection, segmentation, and morphology decrease by 3.76, 3.29, and 2.26 points, while VQA and MLM decrease by 0.64 and 0.52 points, and diagnosis decreases by 3.19 points. Because this variant retains contextual event embeddings but replaces the refined class, box, and morphology attributes with their frozen anchor values, the result suggests that residual refinement primarily improves the spatial and morphological fidelity of local evidence. Retaining contextual event embeddings limits the reductions in language performance, while the diagnosis difference suggests that refined cellular attributes also contribute to patient-level prediction.

Removing zero gating reduces detection, morphology, and diagnosis by 1.31, 1.13, and 3.06 points, respectively. This pattern is consistent with zero gating acting mainly as an optimization stabilizer: it preserves the anchor solution at initialization and allows residual corrections to be introduced gradually rather than forcing immediate departure from the pretrained cellular hypotheses. Taken together, the task-selective degradation patterns indicate a complementary division of labor: residual refinement improves the quality of explicit local attributes, sparse experts improve the transferability of contextual event features, and zero gating regulates how safely the former is adapted from the frozen anchor.

Patient evidence aggregation. Table 3 compares alternative ways of organizing cellular records for diagnosis. Mean pooling can dilute abnormalities represented by few cells, whereas max pooling emphasizes extreme responses without explicitly retaining their prevalence. The View Transformer achieves 65.12 balanced accuracy, compared with 59.40 for mean pooling and 56.30 for max pooling. Full HPEM reaches 75.01, exceeding the View Transformer by 9.89 points. This pattern supports combining cross-view modeling with explicit cellular-composition and morphology summaries.

Table 3: HPEM ablation on the held-out 68-patient test set (mean$_{\text{std}}$, %).

Variant	Acc.	Bal. Acc.	Macro-F1
Alternative aggregation			
Mean pooling	$58.82_{1.47}$	$59.40_{1.62}$	$56.80_{1.75}$
Max pooling	$55.88_{1.47}$	$56.30_{1.88}$	$54.10_{1.92}$
View Transformer	$64.71_{2.55}$	$65.12_{1.51}$	$63.05_{1.58}$
HPEM component ablations			
w/o sparse-evidence attention	$70.59_{1.47}$	$70.92_{1.38}$	$68.85_{1.42}$
w/o expected counts	$69.12_{2.55}$	$69.45_{0.96}$	$67.20_{1.05}$
w/o morphology moments	$64.71_{2.94}$	$65.10_{1.54}$	$63.15_{1.60}$
w/o View Transformer	$60.29_{1.47}$	$60.70_{1.56}$	$58.50_{1.65}$
Full Cell2Case	$\mathbf{74.02}_{4.73}$	$\mathbf{75.01}_{3.39}$	$\mathbf{72.26}_{4.39}$

Among the reported component variants, removing the View Transformer corresponds to the largest balanced-accuracy difference of 14.31 points. Removing morphology moments, expected counts,

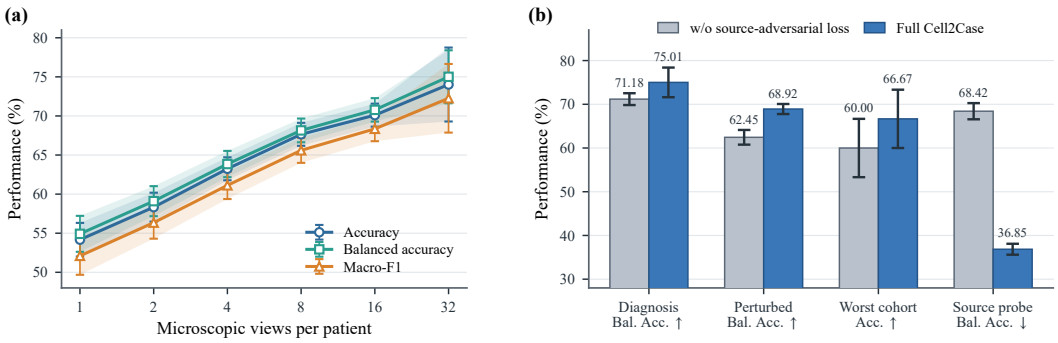


Figure 3: Evidence accumulation and source dependence. **(a)** Diagnosis performance across view budgets, averaged over 10 subsamplings for each of three seeds. **(b)** Effects of source-adversarial learning on diagnosis, perturbation robustness, worst-cohort accuracy, and source prediction. Shading and error bars show one standard deviation; panel (b) uses a truncated vertical axis.

and sparse-evidence attention corresponds to differences of 9.91, 5.56, and 4.09 points, respectively. These comparisons support the complementary value of cross-view relationships, morphological variation, cellular composition, and confidence-weighted event summaries. HPEM therefore organizes evidence beyond a single pooled description of the patient.

Evidence accumulation and source dependence. We examine how diagnosis performance changes with the view budget. Figure 3(a) reports increasing balanced accuracy from 54.92 at one view to 75.01 at 32 views, alongside an increase in macro-F1 from 52.10 to 72.26. The reported trend supports the value of incorporating observations from multiple patient-associated views, suggesting that diagnostic evidence is distributed across different microscopic fields.

Figure 3(b) reports balanced accuracy of 75.01 for Full Cell2Case and 71.18 without source-adversarial learning. The reported perturbed-input and worst-cohort scores are 68.92 versus 62.45 and 66.67 versus 60.00, respectively, while source-probe balanced accuracy is 36.85 versus 68.42. The larger gains under perturbation and for the worst-performing cohort suggest that source-adversarial learning mainly improves robustness to cohort-related variation. The reduced source-probe accuracy is further consistent with less linearly recoverable cohort information. These results concern the pooled-cohort task rather than unseen-cohort generalization.

5 Conclusion and Limitations

Cell2Case connects cell-level perception with patient-level diagnosis by preserving query-aligned predictions as structured cellular evidence and hierarchically aggregating them across patient-associated microscopic views. Under the shared evaluation protocol, Cell2Case achieves the highest run-averaged score among the evaluated baselines on every supported task, while multi-task training improves all six non-detection tasks and maintains detection performance. The ablation and view-scaling results further suggest that effective unification benefits from both high-quality cellular evidence and an aggregation mechanism that preserves salient cellular events, expected cell-type counts, morphology statistics, and cross-view interactions.

The patient-level evaluation nevertheless has two main limitations. It uses a modest 68-patient test set, with several diagnostic classes represented by only a few patients, and its seven-class label space is not fully crossed across the three source cohorts. The reported results therefore establish performance on patient-disjoint pooled-cohort classification, but do not demonstrate generalization to an unseen institution. Future work should evaluate Cell2Case on larger, independently collected cohorts in which the complete diagnostic taxonomy is represented within each cohort, and examine how additional clinical and laboratory evidence can be incorporated into the cell-to-case hierarchy.

AI USE DISCLOSURE

Generative AI tools were used to support literature retrieval and discovery, drafting portions of the manuscript, and language editing for clarity and readability. All AI-assisted literature suggestions, citations, and drafted or edited text were independently reviewed and verified by the authors against the primary sources and manuscript content. Generative AI was not used to generate synthetic experimental datasets or establish mathematical claims. The authors take full responsibility for the final manuscript.

REFERENCES

Syed Saiden Abbas and Tjeerd M. H. Dijkstra. Malaria-Detection-2019, 2019. URL https://doi.org/10.17632/5bf2kmwvfn.1.

Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474, 2020. doi: 10.1016/j.dib.2020.105474.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909.

Alexandra Bodzas, Pavel Kodytek, and Jan Zidek. A high-resolution large-scale dataset of pathological and normal white blood cells. *Scientific Data*, 10(1):466, 2023.

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. doi: 10.1038/s41591-019-0508-1.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, abs/2102.04306, 2021. URL https://arxiv.org/abs/2102.04306.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part V*, volume 13435 of *Lecture Notes in Computer Science*, pp. 679–689. Springer, 2022. doi: 10.1007/978-3-031-16443-9_65. URL https://doi.org/10.1007/978-3-031-16443-9_65.

Muhammed Furkan Dasdelen, Ivan Kukuljan, Peter Lienemann, Fatih Ozlugedik, Ario Sadafi, Matthias Hehr, Karsten Spiekermann, Christian Pohlkamp, and Carsten Marr. Ai-based hematological malignancy prediction from peripheral blood smears in a large diagnostic laboratory cohort. *Leukemia*, 40:1318–1322, 2026. doi: 10.1038/s41375-026-02934-1.

José Guilherme de Almeida, Emma Gudgin, Martin Besser, William G. Dunn, Jonathan Cooper, Torsten Haferlach, George S. Vassiliou, and Moritz Gerstung. Computational analysis of peripheral blood smears detects disease-associated cytomorphologies. *Nature Communications*, 14: 4378, 2023. doi: 10.1038/s41467-023-39676-y.

Mohamed Elmanna, Ahmed Elsafty, Yomna Ahmed, Muhammad Ali Rushdi, and Ahmed Morsy. Deep learning segmentation and classification of red blood cells using a large multi-scanner dataset. *CoRR*, abs/2403.18468, 2024. doi: 10.48550/ARXIV.2403.18468. URL https://doi.org/10.48550/arXiv.2403.18468.

Lu Gan and Xi Li. TXL-PBC: a freely accessible labeled peripheral blood cell dataset. *CoRR*, abs/2407.13214, 2024. doi: 10.48550/ARXIV.2407.13214. URL https://doi.org/10.48550/arXiv.2407.13214.

Manuel González-Hidalgo, F. A. Guerrero-Peña, Silena Herold-García, Antoni Jaume-i Capó, and Pedro D. Marrero-Fernández. Red blood cell cluster separation from digital images for use in sickle cell disease. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1514–1525, 2015. doi: 10.1109/JBHI.2014.2356402.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Matthias Hehr, Ario Sadafi, Christian Matek, Peter Lienemann, Christian Pohlkamp, Torsten Haferlach, Karsten Spiekermann, and Carsten Marr. Explainable ai identifies diagnostic cells of genetic aml subtypes. *PLOS Digital Health*, 2(3):e0000187, 2023. doi: 10.1371/journal.pdig.0000187.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29(9):2307–2316, 2023. doi: 10.1038/s41591-023-02504-3.

Jane Hung and Anne E. Carpenter. Applying faster R-CNN for object detection on malaria images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 808–813. IEEE Computer Society, 2017. doi: 10.1109/CVPRW.2017.112. URL https://doi.org/10.1109/CVPRW.2017.112.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2132–2141. PMLR, 2018. URL http://proceedings.mlr.press/v80/ilse18a.html.

Valentin Koch, Sophia J. Wagner, Salome Kazeminia, Ece Sancar, Matthias Hehr, Julia A. Schnabel, Tingying Peng, and Carsten Marr. Dinobloom: A foundation model for generalizable cell embeddings in hematology. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part XII*, volume 15012 of *Lecture Notes in Computer Science*, pp. 520–530. Springer, 2024. doi: 10.1007/978-3-031-72390-2_49. URL https://doi.org/10.1007/978-3-031-72390-2_49.

Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir M. Rajpoot. Nuclick: A deep learning framework for interactive segmentation of microscopic images. *Medical Image Anal.*, 65:101771, 2020. doi: 10.1016/J.MEDIA.2020.101771. URL https://doi.org/10.1016/j.media.2020.101771.

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): 1123, 2022. doi: 10.1038/s41598-021-04426-x.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Andrea Loddo, Cecilia Di Ruberto, Michel Kocher, and Guy Prod'Hom. MP-IDB: the malaria parasite image database for image processing and analysis. In Natasha Leporé, Jorge Brieva, Eduardo Romero, Daniel Racoceanu, and Leo Joskowicz (eds.), *Processing and Analysis of Biomedical Information - First International SIPAIM Workshop, SaMBa 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Revised Selected Papers*, volume 11379 of *Lecture Notes in Computer Science*, pp. 57–65. Springer, 2018. doi: 10.1007/978-3-030-13835-6_7. URL https://doi.org/10.1007/978-3-030-13835-6_7.

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

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

Petru Manescu, Priya Narayanan, Christopher Bendkowski, Muna Elmi, Remy Claveau, Vijay Pawar, Biobele J. Brown, Mike Shaw, Anupama Rao, and Delmiro Fernandez-Reyes. Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning. *Scientific Reports*, 13(1):2562, 2023. doi: 10.1038/s41598-023-29160-4.

Christian Matek, Simone Schwarz, Karsten Spiekermann, and Carsten Marr. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nat. Mach. Intell.*, 1(11):538–544, 2019. doi: 10.1038/S42256-019-0101-9. URL https://doi.org/10.1038/s42256-019-0101-9.

Christian Matek, Sebastian Krappe, Christian Münzenmayer, Torsten Haferlach, and Carsten Marr. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood*, 138(20):1917–1927, 2021. doi: 10.1182/blood.2020010568.

S Mourya, S Kant, P Kumar, A Gupta, and R Gupta. All challenge dataset of isbi 2019 (c-nmc 2019)(version 1)[dataset]. *The Cancer Imaging Archive*, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Jimut Bahan Pal, Aniket Bhattacharyea, Debasis Banerjee, and Br Tamal Maharaj. Advancing instance segmentation and wbc classification in peripheral blood smear through domain adaptation: A study on pbc and the novel rv-pbs datasets. *Expert Systems with Applications*, 249:123660, 2024.

Abdul Rehman, Talha Meraj, Aiman Mahmood Minhas, Ayisha Imran, Mohsen Ali, and Waqas Sultani. A large-scale multi domain leukemia dataset for the white blood cells detection with morphological attributes for explainability. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part III*, volume 15003 of *Lecture Notes in Computer Science*, pp. 553–563. Springer, 2024. doi: 10.1007/978-3-031-72384-1_52. URL https://doi.org/10.1007/978-3-031-72384-1_52.

Abdul Rehman, Iqra Rasool, Ayisha Imran, Mohsen Ali, and Waqas Sultani. Uni-hema: Unified model for digital hematopathology. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 37578–37589. Computer Vision Foundation, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Rehman_Uni-Hema_Unified_Model_for_Digital_Hematopathology_CVPR_2026_paper.html.

Seyed Hamid Rezatofighi and Hamid Soltanian-Zadeh. Automatic recognition of five types of white blood cells in peripheral blood. *Comput. Medical Imaging Graph.*, 35(4):333–343, 2011. doi: 10.1016/J.COMPMEDIMAG.2011.01.003. URL https://doi.org/10.1016/j.compmedimag.2011.01.003.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells

III, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

Muhammad Shahzad, Syed Hamad Shirazi, Muhammad Yaqoob, Zakir Khan, Assad Rasheed, Israr Ahmed Sheikh, Asad Hayat, and Huiyu Zhou. Anerbc dataset: a benchmark dataset for computer-aided anemia diagnosis using RBC images. *Database J. Biol. Databases Curation*, 2024, 2024. doi: 10.1093/DATABASE/BAAE120. URL https://doi.org/10.1093/database/baae120.

Usman Ali Shams, Isma Javed, Muhammad Fizan, Aqib Raza Shah, Ghulam Mustafa, Muhammad Zubair, Yehia Massoud, Muhammad Qasim Mehmood, and Muhammad Asif Naveed. Bionet dataset: Ai-based diagnostic solutions using peripheral blood smear images. *Blood Cells, Molecules, and Diseases*, 105:102823, 2024. doi: 10.1016/j.bcmd.2024.102823.

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 Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 2136–2147, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html.

John-William Sidhom, Ingharan J. Siddarthan, Bo-Shiun Lai, Adam Luo, Bryan C. Hambley, Jennifer Bynum, Amy S. Duffield, Michael B. Streiff, Alison R. Moliterno, Philip Imus, Christian B. Gocke, Lukasz P. Gondek, Amy E. DeZern, Alexander S. Baras, Thomas Kickler, Mark J. Levis, and Eugene Shenderov. Deep learning for diagnosis of acute promyelocytic leukemia via recognition of genomically imprinted morphologic features. *npj Precision Oncology*, 5:38, 2021. doi: 10.1038/s41698-021-00179-y.

Siddharth Singi, Shenghuan Sun, Zhanghan Yin, Riya Gupta, Dylan C. Webb, Khawaja H. Bilal, Deepika Dilip, Linlin Wang, Neeraj Kumar, Swaraj Nanda, et al. Interpretable multiple instance learning for hematologic diagnosis from peripheral blood smears. *Communications Medicine*, 6: 309, 2026. doi: 10.1038/s43856-026-01558-x.

Waqas Sultani, Wajahat Nawaz, Syed Javed, Muhammad Sohail Danish, Asma Saadia, and Mohsen Ali. Towards low-cost and efficient malaria detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 20655–20664. IEEE, 2022. doi: 10.1109/CVPR52688.2022.02003. URL https://doi.org/10.1109/CVPR52688.2022.02003.

Shenghuan Sun, Zhanghan Yin, Jacob G Van Cleave, Linlin Wang, Brenda Fried, Khawaja H Bilal, Fabienne Lucas, Irem S. Isgor, Dylan C Webb, Siddharth Singi, et al. Deepheme, a high-performance, generalizable deep ensemble for bone marrow morphometry and hematologic diagnosis. *Science translational medicine*, 17(802):eadq2162, 2025.

Haval Taha, Fattah Alizadeh, and Nawsherwan Mohammad. Creating a white blood cell dataset for segmentation, 2023. URL https://doi.org/10.17632/jzdj6h7gms.2.

Satoshi Tsutsui, Winnie Pang, and Bihan Wen. Wbcatt: A white blood cell dataset annotated with detailed morphological attributes. 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/9f34484e5b8d87f09cc58c292a1c9f5d-Abstract-Datasets_and_Benchmarks.html.

Feng Yang, Mahdieh Poostchi, Hang Yu, Zhou Zhou, Kamolrat Silamut, Jian Yu, Richard James Maude, Stefan Jäger, and Sameer K. Antani. Deep learning for smartphone-based malaria parasite detection in thick blood smears. *IEEE J. Biomed. Health Informatics*, 24(5):1427–1438, 2020. doi: 10.1109/JBHI.2019.2939121. URL https://doi.org/10.1109/JBHI.2019.2939121.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10:41, 2023. doi: 10.1038/s41597-022-01721-8.

Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 16965–16974. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01605. URL https://doi.org/10.1109/CVPR52733.2024.01605.

A Appendix

A.1 Dataset Details

The cellular and language benchmark comprises 46 dataset configurations drawn from 36 public source collections. These configurations comprise 16 classification, 17 detection, 11 segmentation, one VQA, and one MLM configuration. Morphology supervision is attached to LeukemiaAttri detection instances and is therefore not counted as a separate dataset configuration. A source collection reused across tasks contributes to each applicable task, so image counts across rows in Table 4 are not additive. Dataset-specific labels are mapped to shared task taxonomies before training.

The classification and detection taxonomies cover mature and immature blood cells, erythroid cells, platelets, artifacts, blood-borne parasites, and dataset-specific difficult or unidentified instances. Morphology supervision contains six binary attributes: nuclear chromatin, nuclear shape, nucleolus, cytoplasm, cytoplasmic basophilia, and cytoplasmic vacuoles. Each attribute has 32,117 valid labels, with 10,545, 30,173, 14,687, 21,534, 20,543, and 1,974 positive labels, respectively. Unknown attributes are excluded through per-attribute validity masks.

Segmentation sources cover leukocytes, erythrocytes, and malaria parasites. Their masks are converted to a common foreground convention, and multiple masks associated with the same image are merged by union. VQA contains yes/no, count, cell-type, morphology, and open-ended questions derived from WBCAtt (Tsutsui et al., 2023). The MLM subset contains masked descriptions derived from LeukemiaAttri (Rehman et al., 2024). VQA and MLM are evaluated as distinct subsets.

A.2 Patient Cohorts and Data Partitions

Patient-level diagnosis combines JHU APL/AML (Sidhom et al., 2021), MLL Helmholtz AML-Cytomorphology (Hehr et al., 2023), and the labeled portion of C-NMC 2019 (Mourya et al., 2019). These cohorts contain 396 patients assigned to seven harmonized diagnostic classes, as summarized in Table 5. The unlabeled final phase of C-NMC 2019 is excluded.

JHU contributes PML–RARA and AML-unspecified cases, MLL contributes Control, PML–RARA, NPM1, CBFB–MYH11, and RUNX1–RUNX1T1 cases, and C-NMC contributes Control and B-ALL cases. Consequently, diagnostic label support is not fully crossed across cohorts. The 54 diagnosis-validation and 68 diagnosis-test patients are reserved before constructing any task-specific partition. All images, cell crops, annotations, and language records associated with these patients are excluded from every model-training partition. Sources with uncertain patient-level provenance were excluded to prevent overlap between the reserved diagnosis patients and model-training data.

A.3 Experimental Settings

Architecture details. For each image, the anchor provides decoder queries, class logits, normalized boxes, and morphology logits, forming $(\mathbf{Q}_0, \mathbf{C}_0, \mathbf{B}_0, \mathbf{M}_0)$. At most 300 queries are retained according to maximum sigmoid class confidence. These anchor outputs define the initial cellular hypotheses for residual refinement, and gradients from Cell2Case are never propagated into the Uni-Hema anchor. The visual pathway uses frozen DINOv2 ViT-S/14 and a four-block convolutional stream producing 192-dimensional features at strides 4, 8, and 16. Each task adapter is a zero-initialized $192 \to 48 \to 192$ residual bottleneck. The residual refiner uses eight attention heads; only the final two BioClinicalBERT layers are updated (Alsentzer et al., 2019). The two-layer View Transformer also uses eight heads. Patient-level aggregation uses at most 32 views.

Table 4: Composition of the cellular and language benchmark. "Configs." denotes dataset configurations. Morphology supervision is attached to LeukemiaAttri detection instances and is not counted separately; images reused across tasks are counted in each relevant row.

Task	Configs	Source collections	Images	Annots.	Targets
Classification	16	BMC; AML Matek; Raabin-WBC; Warty Pig; LISC; KRD-WBC; BCCD; HRLS; APL/AML; White-Blood-Cell Dataset; Acevedo-PBC; RV-PBS; PBC-8-DA; C-NMC 2019; BloodMNIST; AML Hehr	226,530	226,544	45 cell and parasite classes
Detection	17	LeukemiaAttri; M5; TXL-PBC; BCCD; Sickle-cell; Plasmodium; Plasmodium Phonecamera; Tuberculosis Phonecamera; Vivax; Thick Blood Smears; NIH–NLM Thick PV; Parasite; Acevedo-PBC; MP-IDB; Bio-Net; Raabin-M1; Raabin-M2	43,943	387,945	29 foreground classes plus no-object
Morphology	–	Attributes attached to LeukemiaAttri detection instances	32,117 cells; 192,702 valid labels		Six binary attributes with validity masks
Segmentation	11	Malaria-Detection-2019; NuClick; KRD-WBC; WBC Image Dataset; Raabin-WBC; ErythrocytesIDB; AneRBC-II Anemic; AneRBC-II Healthy; MP-IDB; Elsafty-RBC; BBBC041Seg	205,750	205,830	Binary foreground masks
VQA	1	WBCAtt-VQA	3,073	12,339	Five question types and free-text answers
MLM	1	LeukemiaAttri-MLM	1,607	5,552	Masked hematology descriptions

Sources. Classification: (Matek et al., 2021; 2019; Kouzehkanan et al., 2022; Rezatofighi & Soltanian-Zadeh, 2011; Taha et al., 2023; Bodzas et al., 2023; Sidhom et al., 2021; Acevedo et al., 2020; Pal et al., 2024; Mourya et al., 2019; Yang et al., 2023; Hehr et al., 2023). Detection: (Rehman et al., 2024; Sultani et al., 2022; Gan & Li, 2024; Hung & Carpenter, 2017; Yang et al., 2020; Acevedo et al., 2020; Loddo et al., 2018; Shams et al., 2024; Kouzehkanan et al., 2022). Morphology: (Rehman et al., 2024). Segmentation: (Abbas & Dijkstra, 2019; Koohbanani et al., 2020; Taha et al., 2023; González-Hidalgo et al., 2015; Kouzehkanan et al., 2022; Shahzad et al., 2024; Loddo et al., 2018; Elmanna et al., 2024; Hung & Carpenter, 2017). VQA and MLM: (Tsutsui et al., 2023; Rehman et al., 2024).

Segmentation-ontology conditioning. Each segmentation record is associated with an ontology identifier $d_s \in \{0, \dots, 5\}$ defined by the source dataset metadata. It is not inferred from the image and is unrelated to the patient-cohort source label used for source-adversarial diagnosis training. A learned embedding produces channel-wise scale and shift vectors,

$$[\boldsymbol{\gamma}_{d_s}; \boldsymbol{\beta}_{d_s}] = \mathrm{Emb}(d_s), \qquad \widetilde{\mathbf{P}}_4 = \phi_p(\mathbf{P}_4) \odot (1 + \boldsymbol{\gamma}_{d_s}) + \boldsymbol{\beta}_{d_s}, \tag{9}$$

which condition the pixel representation before query–pixel mask prediction.

Progressive optimization. Table 6 summarizes the trainable and fixed modules in each stage. The detector anchor and DINOv2 encoder remain frozen throughout training.

Training details. Task steps are determined by the prescribed epoch exposure and randomly interleaved within each stage. Detection, classification, and segmentation use one epoch; VQA and MLM use three epochs; and diagnosis uses 60 epochs. Stage IV adds 0.5 epoch for each language task and 12 epochs for diagnosis. For patients with more than 32 views, training uses a rolling window over a fixed per-patient shuffled order, whereas validation and test use 32 deterministically spaced views.

Table 5: Patient-level dataset composition. Patients are disjoint across diagnosis splits; validation and test patients are also excluded from every model-training partition across all tasks.

(a) Cohort composition

Cohort	Images	Train	Val.	Test	Available labels
JHU APL/AML	25,724	70	12	24	PML–RARA, AML-unspecified Control, PML–RARA, NPM1,
MLL Helmholtz	81,214	131	29	29	CBFB–MYH11, RUNX1–RUNX1T1
C-NMC 2019	12,528	73	13	15	Control, B-ALL
Total	119,466	274	54	68	Seven classes

(b) Diagnosis distribution

Diagnosis	Train	Val.	Test	Total
Control	68	16	17	101
PML–RARA	35	7	16	58
NPM1	26	5	5	36
CBFB–MYH11	25	6	6	37
RUNX1–RUNX1T1	22	5	5	32
AML-unspecified	51	9	12	72
B-ALL	47	6	7	60
Total	274	54	68	396

Table 6: Progressive training schedule. Modules not listed as trainable in a stage remain fixed.

Stage	Trainable modules	Fixed modules	Objectives	LR
I	Multi-scale visual stream, gated fusion, detection adapter, and residual refiner	Detector anchor, DINOv2, and remaining task modules	Det., Morph.	3×10^{-5}
II	Multi-scale visual stream, gated fusion, classification and segmentation adapters, classifier, mask decoder, and sparse experts	Detector anchor, DINOv2, residual refiner, language module, and HPEM	Cls., Seg.	5×10^{-5}
III	Language and case adapters, sparse experts, event–language module (last two BERT layers trainable), HPEM, and diagnosis head	Detector anchor, DINOv2, visual stream, fusion, refiner, classifier, and mask decoder	VQA, MLM, Diag.	2×10^{-5}
IV	Language and case adapters, event–language module (last two BERT layers trainable), HPEM, and diagnosis head	Detector anchor, DINOv2, visual modules, and sparse experts	VQA, MLM, Diag. calibration	1×10^{-5}

Baseline implementation. All baselines are evaluated on the same fixed, training-disjoint partitions used for Cell2Case, with predictions mapped to the same task taxonomies and evaluated using the same task-level metrics. Published results obtained on different partitions are not directly transferred into our comparison. Method-specific input resolution and preprocessing are retained when required.

For detection, RT-DETR uses a PResNet-18 backbone with 352×352 inputs and three decoder layers, while AttriDet follows its released detection configuration. Classification baselines use 224×224 inputs: ResNet-50 and DinoBloom-S provide pretrained visual representations followed by task-specific linear classification heads, with the pretrained encoders kept frozen during classifier training. U-Net and TransUNet are trained on the shared binary segmentation partitions with 336×336 inputs. For language tasks, M^3AE is fine-tuned from its released pretrained initialization on the corresponding VQA/MLM partitions, whereas BLIP-2 is evaluated with its supported zero-shot VQA inference.

For patient-level comparison, ABMIL, TransMIL, MILLIE, CAREMIL, and cAItomorph use the same frozen DINOv2-S view representations, patient manifests, and seven-class diagnostic taxonomy. Each patient is represented by at most 32 view features, matching the Cell2Case view budget; training and validation/test view selection follow the same patient-level sampling protocol. Patient-level model selection uses only the diagnosis-validation set and never accesses the held-out test set. Thus, comparisons control data partitions, target definitions, representation input, and evaluation metrics while retaining each baseline's aggregation architecture.

Inference settings. At inference, a fixed query-level safety gate controls the refined cellular attributes. For each query, the refined class, box, and morphology predictions are retained only when the refined class confidence exceeds the anchor confidence by at least 0.02 and the maximum normalized box displacement is at most 0.12; otherwise, the corresponding anchor attributes are used. The contextual event embedding remains unchanged. CBME-dependent downstream branches therefore consume the gated cellular attributes together with the refined event embedding. Detection uses a score floor of 0.001 and an NMS IoU threshold of 0.5 for mAP_{50} evaluation. Segmentation probabilities are thresholded at 0.5. Classification, morphology, MLM, and diagnosis use argmax prediction. Open-ended VQA answers are decoded greedily for at most 64 tokens.

Metric implementation. Morphology is evaluated after Hungarian matching. Only matched cells with box IoU ≥ 0.5 contribute to attribute evaluation. For each binary morphology attribute a, we compute

$$F_{1,a} = \frac{2TP_a}{2TP_a + FP_a + FN_a}. \tag{10}$$

where only valid attribute annotations are included. Let

$$\rho = \frac{N_{\text{valid,matched}}}{N_{\text{valid,GT}}} \tag{11}$$

denote the fraction of valid ground-truth attribute annotations belonging to successfully matched cells. We report coverage-adjusted macro-F1 as

$$\text{CA-MacroF1} = \rho \frac{1}{|\mathcal{A}'|} \sum_{a \in \mathcal{A}'} F_{1,a}, \tag{12}$$

where $\mathcal{A}'$ contains attributes with valid matched annotations. The coverage term penalizes morphology predictions that are accurate only for the subset of successfully localized cells.

For VQA, each method first produces its final textual answer using its native inference mechanism. Cell2Case uses question-type-aware hybrid inference: predicted yes/no, count, cell-type, and morphology questions use their corresponding structured prediction heads, whereas open-ended questions use autoregressive generation. All methods are then evaluated using the same answer-level token-F1. Reference and predicted answers are lower-cased and tokenized into alphanumeric tokens. Let T and $\widehat{T}$ denote the resulting token multisets and let $o = |T \cap \widehat{T}|$ denote their multiset overlap. We compute

$$P = \frac{o}{|\widehat{T}|}, \qquad R = \frac{o}{|T|}, \qquad F_1 = \frac{2PR}{P+R}. \tag{13}$$

The reported Answer Token-F1 is the arithmetic mean over all VQA records. Cell2Case routing uses the predicted rather than ground-truth question type, so routing errors are reflected in the final answer score.

MLM accuracy is computed only at masked positions, and diagnosis metrics are computed per patient. Single-task variants follow the same fixed-schedule protocol and use the same initialization, data partitions, model capacity, and task-specific training exposure as the corresponding branches of the multi-task model.

Aggregation across datasets. For each task, predictions from all associated dataset configurations are pooled under the shared taxonomy before computing the reported metric; metrics are not first computed independently and then averaged across configurations. Detection pools all predictions and annotations before computing global-class AP and mAP_{50}. Classification reports mean recall over the 45 global classes. Segmentation accumulates foreground pixel counts over all evaluation images before computing Dice. Morphology pools matched-cell statistics before averaging F1 over the six attributes and applying coverage adjustment. VQA is averaged over question records, whereas MLM accuracy is computed over all masked tokens.

Source-dependence and robustness evaluation. For perturbed-input evaluation, each test view undergoes random color jitter, Gaussian blur, and gamma correction. Brightness and contrast factors are sampled from $[0.8, 1.2]$, the blur-kernel radius from $\{1, 3\}$, and gamma from $[0.85, 1.15]$.

Table 7: Patient-level diagnosis on the held-out 68-patient test set. Results are $\text{mean}_{\text{s.d.}}$ across three training runs; confidence intervals are computed from 100,000 class-stratified bootstrap estimates of the cross-run mean.

Metric	$\text{Mean}_{\text{s.d.}}$	Bootstrap 95% CI
Accuracy	$74.02_{4.73}$	[61.76, 80.88]
Balanced accuracy	$75.01_{3.39}$	[60.92, 83.17]
Macro-F1	$72.26_{4.39}$	[57.45, 81.19]

Perturbations are sampled independently using fixed random seeds while patient membership and view budgets remain unchanged. Results are averaged over five perturbation realizations for each of the three trained models.

Worst-cohort accuracy is the minimum patient-level accuracy among JHU, MLL, and C-NMC. For the source probe, patient representations are frozen and a multinomial linear classifier is trained to predict the three cohort identities. The probe is fitted on diagnosis-training patients, its regularization is selected using diagnosis-validation patients, and balanced accuracy is reported on diagnosis-test patients. Probe fitting does not update Cell2Case.

Algorithm 1 Patient-level inference with Cell2Case

Require: Patient views $\mathcal{X}_p = \{x_{p,f}\}_{f=1}^{F}$
Ensure: Diagnosis probabilities $\hat{\mathbf{y}}_p$
for $f = 1, \ldots, F$ **do**
  $P_f = \{P_{4,f}, P_{8,f}, P_{16,f}\} \leftarrow \text{VisualEncoder}(x_{p,f})$
  $(Q_{0,f}, C_{0,f}, B_{0,f}, M_{0,f}) \leftarrow \text{Anchor}(x_{p,f})$
  $(H_f, C_f, B_f, M_f) \leftarrow \text{ResidualRefiner}(Q_{0,f}, C_{0,f}, B_{0,f}, M_{0,f}, P_f)$
  $E_f \leftarrow \text{SparseExperts}(H_f)$
  $(\widehat{C}_f, \widehat{B}_f, \widehat{M}_f) \leftarrow \text{SafetyGate}(C_f, B_f, M_f; C_{0,f}, B_{0,f}, M_{0,f})$
  $\mathcal{R}_f \leftarrow \{(\widehat{C}_{f,i}, \widehat{B}_{f,i}, \widehat{M}_{f,i}, E_{f,i})\}_{i=1}^{Q}$
  $h_{p,f} \leftarrow \text{EventAgg}(E_f, \widehat{C}_f, \widehat{M}_f)$
end for
$\widetilde{H}_p \leftarrow \text{ViewTransformer}([h_{p,1}, \ldots, h_{p,F}])$
$z_p \leftarrow [\text{Mean}(\widetilde{H}_p); \text{Max}(\widetilde{H}_p); \text{Attn}(\widetilde{H}_p)]$
return $\hat{\mathbf{y}}_p = \text{softmax}(\text{DiagnosisHead}(z_p))$

A.4 Patient-Level Inference

Algorithm 1 summarizes how Cell2Case converts patient-associated microscopic views into a diagnosis. Here, Anchor includes the confidence-based top-Q query selection described in Section 3.2. SafetyGate retains refined class, box, and morphology attributes only when the refinement satisfies the fixed confidence and displacement criteria; otherwise, the corresponding anchor attributes are used, while the contextual event embeddings remain unchanged. EventAgg performs event-to-view aggregation over event embeddings, class logits, and morphology logits using sparse-evidence attention, expected counts, and morphology moments. The resulting view tokens are jointly processed by the View Transformer.

A.5 Additional Patient-Level Analyses

Patient-level uncertainty. We separately quantify variability across training runs and uncertainty from the finite test population. Seed variability is reported as the standard deviation across three runs. For patient-level uncertainty, we generate 100,000 class-stratified bootstrap samples by resampling patients with replacement within each diagnosis class while preserving its original size. The same sampled patient indices are applied to all three runs. Each metric is computed separately for each run and then averaged across runs. The 2.5th and 97.5th percentiles of these run-averaged bootstrap estimates define the 95% confidence interval.

Table 8: Source-only diagnosis baseline on the held-out 68-patient test set. The deterministic source-only baseline predicts the most frequent training diagnosis within each cohort and uses no image-derived information. Cell2Case results are averaged over three independent training runs, with standard deviations shown as subscripts. All values are percentages.

Input	Acc.	Bal. Acc.	Macro-F1
Cohort identity only	41.18	36.13	24.20
Cell2Case	**74.02**$_{4.73}$	**75.01**$_{3.39}$	**72.26**$_{4.39}$

Table 7 shows that the patient-bootstrap intervals remain broad, particularly for macro-F1, whereas variation across training runs is comparatively small. The limited precision therefore mainly reflects the 68-patient test population and the small number of patients in several diagnostic classes. The patient-level results should consequently be interpreted as estimates for the evaluated pooled cohorts rather than precise measures of performance in a broader clinical population.

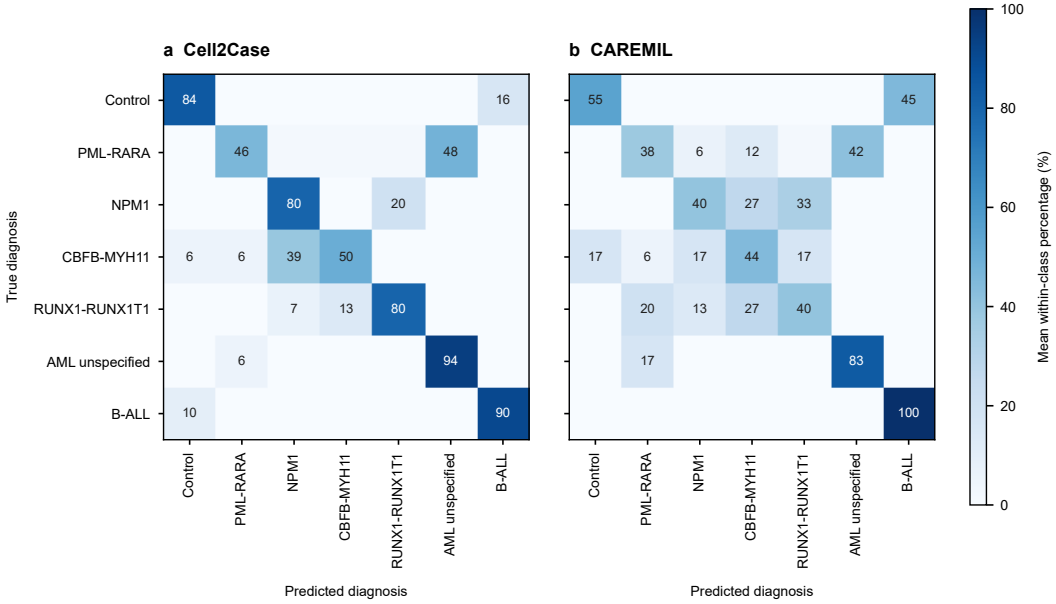


Figure 4: Seven-class patient diagnosis on the held-out 68-patient test set. Row-normalized confusion matrices are averaged across three independent runs for **(a)** Cell2Case and **(b)** CAREMIL. Values are rounded to integer percentages. Numeric annotations below 5% are omitted, while all cells remain represented by color.

Class-wise diagnostic analysis. Figure 4 shows that Cell2Case improves mean recall over CAREMIL for six of the seven classes. The largest gains occur in NPM1 and RUNX1–RUNX1T1, both increasing from 40.00% to 80.00%, followed by Control, increasing from 54.90% to 84.31%. PML–RARA and CBFB–MYH11 remain the most difficult classes for Cell2Case, with recalls of 45.83% and 50.00%. B-ALL is the exception: CAREMIL achieves 100.00% recall compared with 90.48% for Cell2Case, although Cell2Case attains higher B-ALL F1. Thus, the aggregate improvement reflects gains across several classes rather than uniformly better recall. Class-level estimates remain imprecise because several categories contain only five to seven test patients.

Source-only diagnosis baseline. Table 8 evaluates the diagnostic shortcut available from cohort identity alone. For each cohort, the baseline predicts its most frequent diagnosis in the training set: AML-unspecified for JHU, Control for MLL, and B-ALL for C-NMC. Despite using no images, it reaches 41.18% accuracy, confirming the presence of cohort–label dependence. Cell2Case

substantially improves all three metrics, indicating that cohort identity alone cannot explain its performance. Nevertheless, we interpret the task as patient-disjoint pooled-cohort classification rather than cross-cohort generalization.

A.6 Qualitative Results

Qualitative overview across tasks. Figure 5 presents representative outputs across all seven tasks, comparing Cell2Case trained independently for each task (ST), Uni-Hema, and the jointly trained Cell2Case model (MT). For detection in **panel (a)**, both Cell2Case variants recover the annotated blood-cell and platelet instances across TXL-PBC, BCCD, and HCM, while retaining closely matched localization patterns under single- and multi-task training. This behavior is consistent with the anchor-preserving refinement, which allows contextual residual adaptation while limiting disruption of the initial localization hypotheses.

The classification and morphology examples in **panels (b)** and **(c)** illustrate complementary aspects of cellular reasoning. In panel (b), both Cell2Case variants correctly recognize the KRD-WBC basophil and Raabin eosinophil, whereas Uni-Hema misclassifies the displayed examples. Because classification does not directly consume CBME records, the improvement of the MT model is consistent with positive transfer through the shared visual pathway. In panel (c), Cell2Case preserves the lymphoblast identities of the highlighted cells together with their morphology predictions, whereas Uni-Hema assigns the marked cells to neutrophils. The displayed outputs illustrate the benefit of maintaining class and morphology attributes under a common query identity.

Panel (d) compares erythrocyte segmentation on AneRBC-II. All methods recover the overall cellular layout, while the boundary overlay reveals local false-positive and false-negative deviations, particularly in crowded or weakly contrasted regions. Cell2Case MT remains qualitatively close to its ST counterpart and preserves fine-grained boundaries, indicating that joint training does not sacrifice the spatial detail required for dense prediction.

The language examples in **panels (e)** and **(f)** show a clearer effect of shared multi-task supervision. Both Cell2Case variants correctly identify the queried cell as a monocyte in panel (e), whereas Uni-Hema answers "eosinophil." In panel (f), Cell2Case MT reconstructs both the increased lymphoblast burden and the normal neutrophil morphology, matching the reference description. The ST model captures the lymphoblast increase but predicts atypical neutrophil morphology, while Uni-Hema instead predicts a normal WBC count and abnormal morphology. These examples are consistent with the event-grounded language pathway, in which VQA and MLM access the same query-aligned event embeddings used to represent cellular identity, morphology, and local context.

panel (g) illustrates patient-level inference across multiple microscopic views. The displayed C-NMC cases are correctly predicted as B-ALL, while the JHU examples are correctly classified as PML–RARA and AML-unspecified. Although ST and MT reach the same diagnoses in these examples, their per-view attention scores differ, showing that the two training regimes can assign different importance to individual microscopic fields. This behavior is consistent with HPEM, which summarizes each view using salient cellular events, expected cell-type counts, and morphology statistics before modeling dependencies across views. Uni-Hema is not shown in panel (g) because it does not provide patient-level diagnostic outputs.

Additional detection examples. Figures 6– 8 examine three distinct localization regimes. LeukemiaAttri contains leukemic cells embedded among densely distributed erythrocytes, TXL-PBC introduces substantial variation in cell scale and surrounding cellular density, and Plasmodium Phonecamera contains small, weakly contrasted parasite targets. These settings therefore test complementary failure modes, including crowded-cell ambiguity, scale variation, and missed small-object detections. Across the three datasets, Cell2Case ST and MT produce broadly similar localization patterns, supporting the quantitative observation that multi-task training does not destabilize detection. The agreement is particularly notable in dense fields, where additional supervision could otherwise increase redundant responses or shift predictions toward nearby structures. Relative to the task-specific baselines and Uni-Hema, the displayed Cell2Case predictions remain associated with individual cellular structures across a wide range of target sizes. The remaining discrepancies are concentrated in highly crowded regions and very small, low-contrast targets, suggesting that localization uncertainty is driven primarily by weak spatial separation and limited visual evidence rather

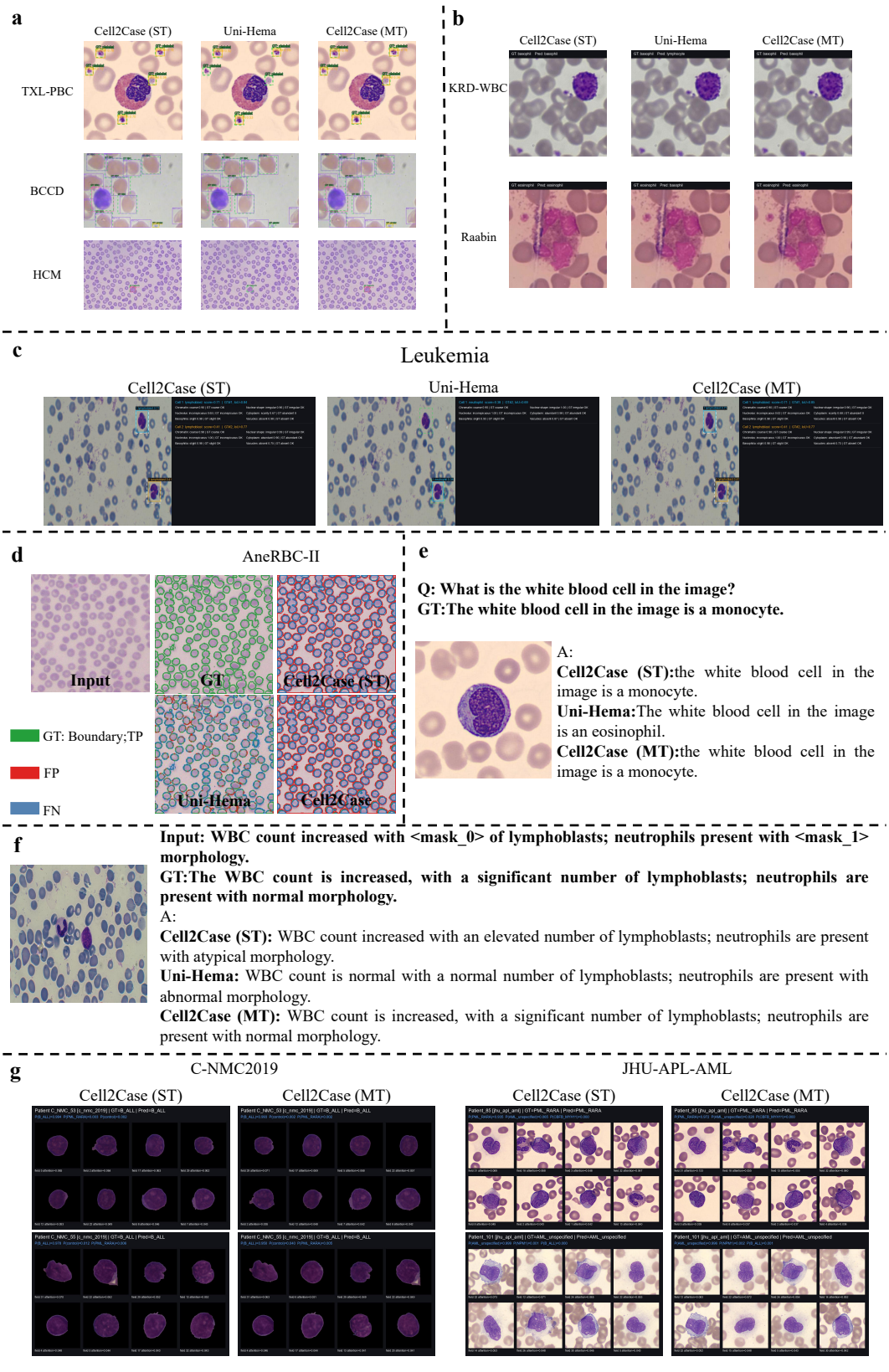


Figure 5: Qualitative overview across seven tasks. Cell2Case with single-task (ST) and multi-task (MT) training is compared with Uni-Hema where supported.

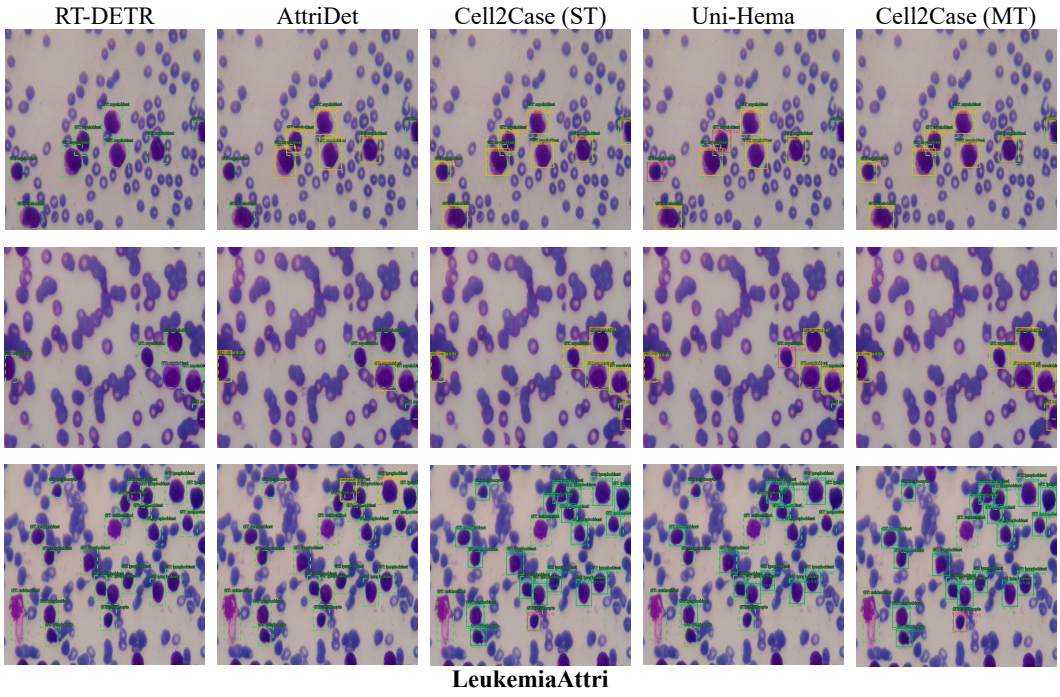


Figure 6: Detection comparison on LeukemiaAttri. Columns show RT-DETR, AttriDet, Cell2Case (ST), Uni-Hema, and Cell2Case (MT), with ground-truth and predicted boxes and labels overlaid.

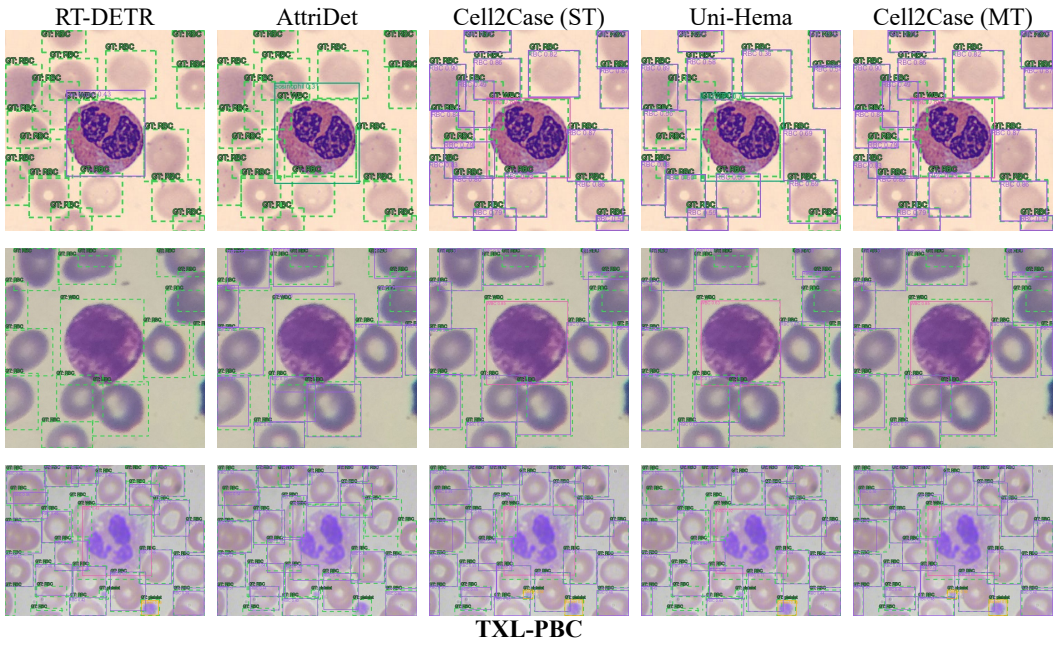


Figure 7: Detection comparison on TXL-PBC across RT-DETR, AttriDet, Cell2Case (ST), Uni-Hema, and Cell2Case (MT).

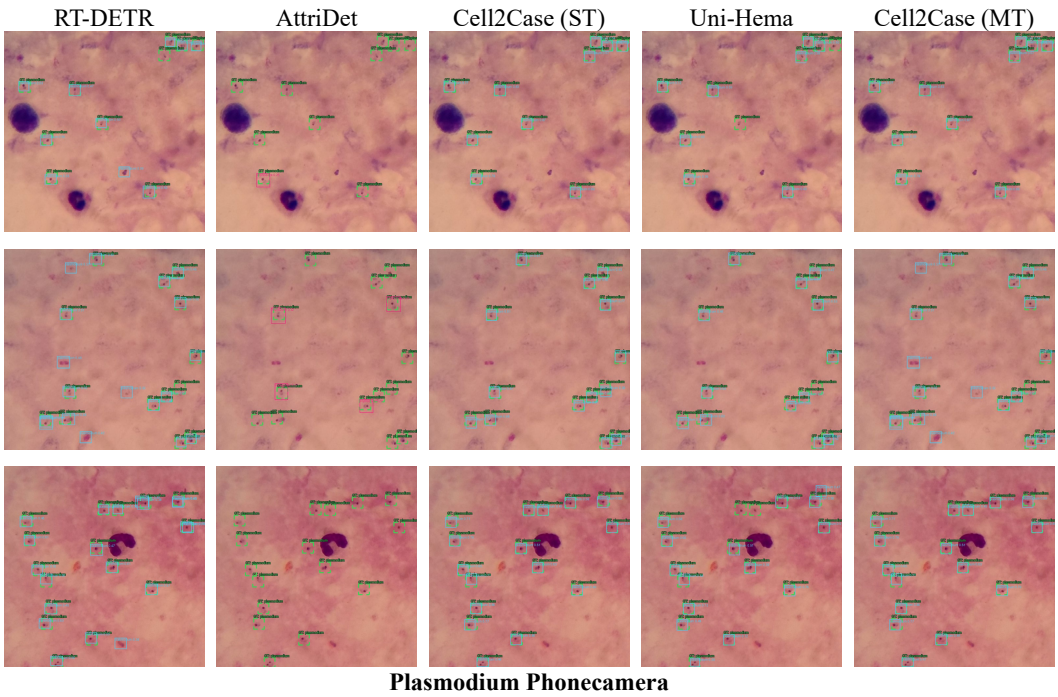


Figure 8: Detection comparison on Plasmodium Phonecamera across RT-DETR, AttriDet, Cell2Case (ST), Uni-Hema, and Cell2Case (MT).

than by multi-task interference. These examples therefore complement the aggregate detection score by showing that the ST–MT stability extends across substantially different microscopic conditions.

Additional classification examples. Figures 9– 11 examine classification under three complementary sources of ambiguity. The BMC examples include closely related bone-marrow cell identities, the Raabin-WBC examples focus on fine-grained granulocyte categories, and the KRD-WBC examples contain prominent erythrocyte or platelet-like background structures that can distract image-level classification. The qualitative differences between ST and MT are not uniform across these settings. On visually distinctive Raabin-WBC granulocytes, both Cell2Case variants recover the correct labels, indicating that joint training preserves already discriminative cellular features. In the more ambiguous BMC examples, MT corrects errors made by ST on fine-grained lymphoid identities, while the KRD-WBC examples show the clearest improvement under background interference, where MT is less frequently attracted to thrombocyte-related categories. Because the classification branch does not directly consume CBME records, these differences are consistent with transfer through the shared visual representation rather than direct reuse of query-level predictions. Taken together, the examples suggest that the main benefit of multi-task training emerges in fine-grained or background-confounded cases, whereas obvious examples remain largely unchanged. This pattern matches the larger quantitative gain observed for classification under multi-task training.

Additional segmentation examples. Figure 12 compares erythrocyte segmentation under both sparse and densely packed cellular distributions. For isolated erythrocytes, most methods recover the overall cell geometry and the remaining differences are dominated by small contour displacements. The crowded examples are more discriminative because touching cells introduce fragmented contours, merged regions, and localized false-positive or false-negative boundaries. Under these harder conditions, Cell2Case MT remains close to its ST counterpart and preserves individual-cell separation in regions where adjacent erythrocytes are difficult to distinguish. The boundary overlay also shows that residual errors are concentrated near touching interfaces and weakly contrasted edges rather than across the complete cell body. This indicates that joint training preserves the fine spatial structure required for dense prediction while still benefiting from shared cellular supervision.

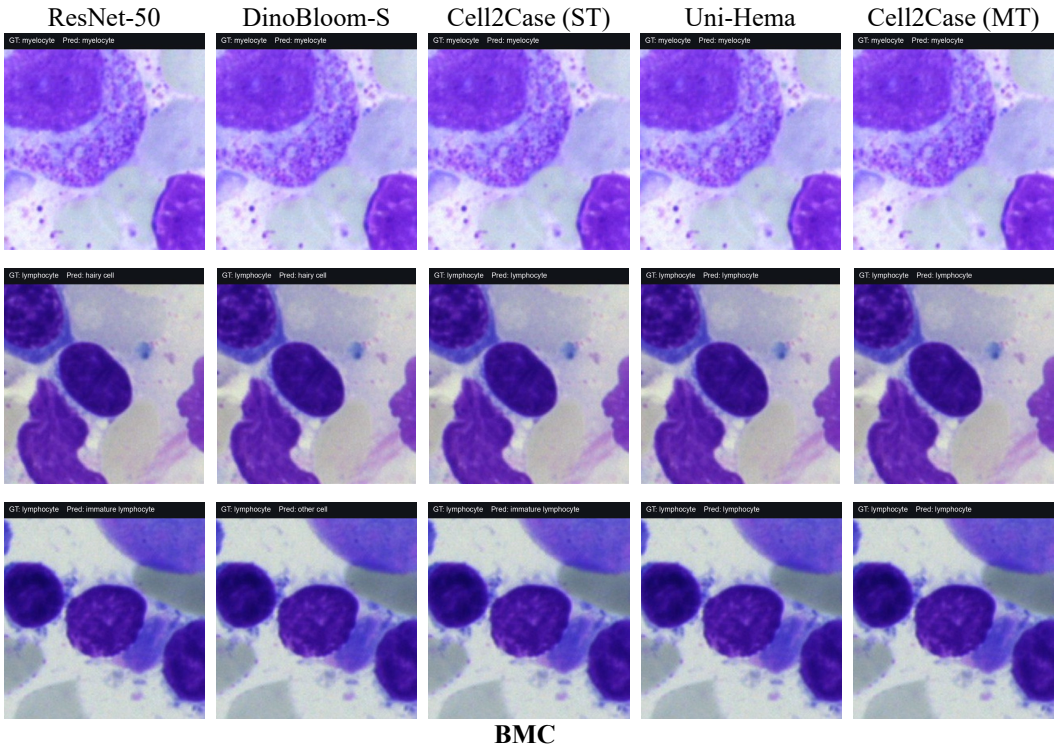


Figure 9: Classification comparison on BMC. Columns show ResNet-50, DinoBloom-S, Cell2Case (ST), Uni-Hema, and Cell2Case (MT); each panel reports the ground-truth and predicted class.

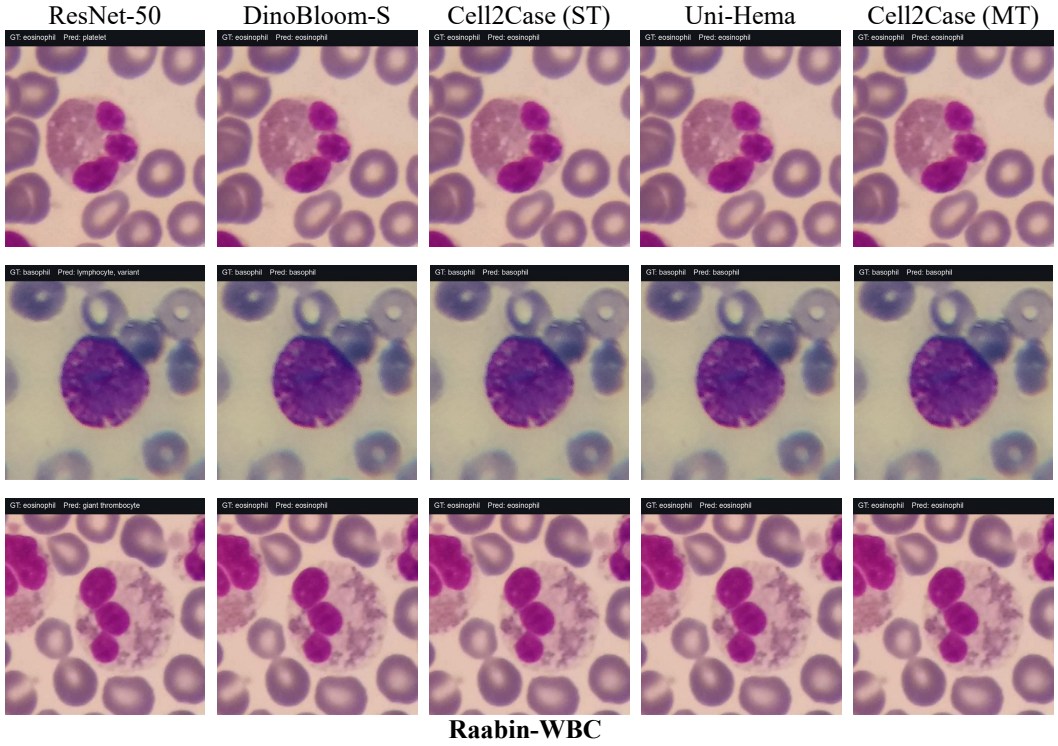


Figure 10: Classification comparison on Raabin-WBC across ResNet-50, DinoBloom-S, Cell2Case (ST), Uni-Hema, and Cell2Case (MT).

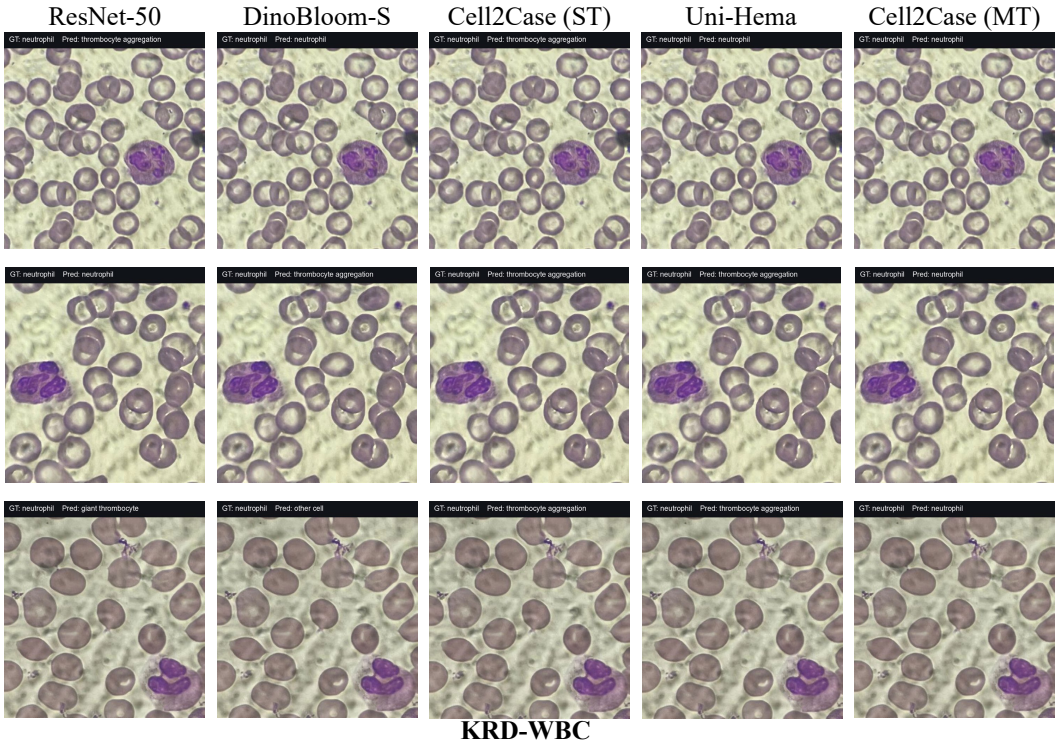


Figure 11: Classification comparison on KRD-WBC across ResNet-50, DinoBloom-S, Cell2Case (ST), Uni-Hema, and Cell2Case (MT).

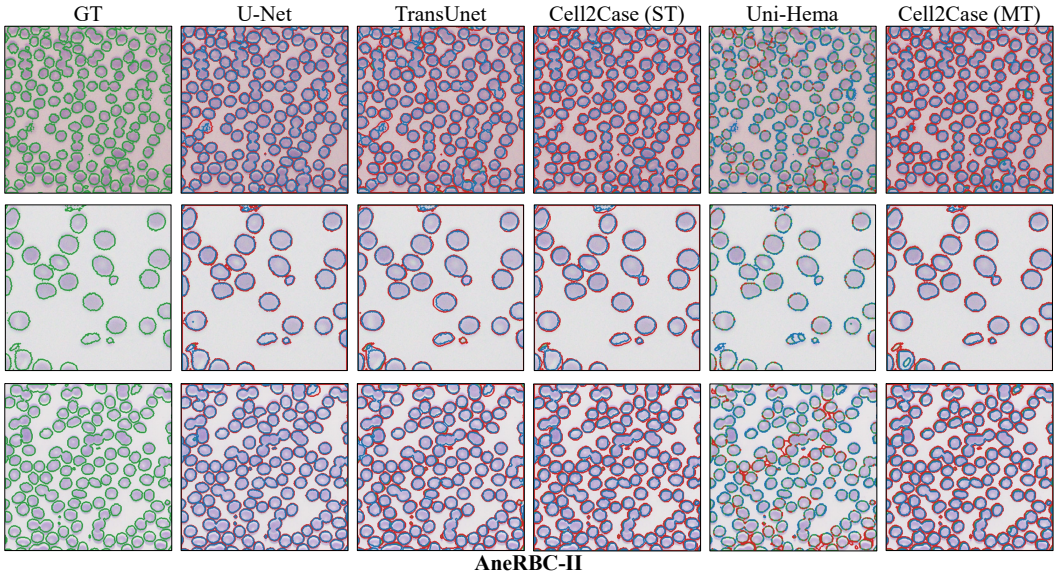


Figure 12: Segmentation comparison on AneRBC-II. Columns show the ground truth, U-Net, TransUNet, Cell2Case (ST), Uni-Hema, and Cell2Case (MT); green, red, and blue denote TP, FP, and FN boundaries.

The qualitative behavior is therefore consistent with the quantitative ST–MT improvement in Dice and suggests that the remaining challenge lies primarily in boundary ambiguity between neighboring cells.