Uni-Hema: Unified Model for Digital Hematopathology

Abdul Rehman[1], Iqra Rasool[2], Ayisha Imran[2], Mohsen Ali[1], Waqas Sultani[1]
[1]Intelligent Machines Lab, Department of Artificial Intelligence, Information Technology University
[2]Hematopathology Department, Chughtai Lab, Lahore
phdcs23002@itu.edu.pk, mohsen.ali@itu.edu.pk, waqas.sultani@itu.edu.pk

Abstract

Digital hematopathology requires cell-level analysis across diverse disease categories, including malignant disorders (e.g., leukemia), infectious conditions (e.g., malaria), and non-malignant red blood cell disorders (e.g., sickle cell disease). Whether single-task, vision-language, WSI-optimized, or single-cell hematology models, these approaches share a key limitation: they cannot provide unified, multi-task, multi-modal reasoning across the complexities of digital hematopathology. To overcome these limitations, we propose ***Uni-Hema****, a multi-task, unified model for digital hematopathology integrating detection, classification, segmentation, morphology prediction, and reasoning across multiple diseases. Uni-Hema leverages 46 publicly available datasets, encompassing over 700K images and 21K question-answer pairs, and is built upon* ***Hema-Former****, a multimodal module that bridges visual and linguistic representations at the hierarchy level for the different tasks (detection, classification, segmentation, morphology, mask language modeling, and visual question answering) at different granularities. Extensive experiments demonstrate that Uni-Hema achieves comparable or superior performance compared to training on single-task and single-dataset models, across diverse hematological tasks, while providing interpretable, morphologically relevant insights at the single-cell level. Our framework establishes a new standard for multi-task and multi-modal digital hematopathology. The code is available at* `https://github.com/intelligentMachines-ITU/Uni-Hema`

1. Introduction

Hematology, the study of blood and its disorders [54], relies heavily on peripheral blood film (PBF) examination for diagnosing conditions such as malaria, anemia, sickle cell disease, thalassemia, and leukemia [6]. Just to give an idea of the scale of its need, malaria caused 263 million cases and 597,000 deaths in 2023, concentrated in the WHO African Region [50], while sickle cell disease (SCD) affected 7.74 million people globally in 2021, resulting in an estimated 376,000 deaths [18] that year. Leukemia further contributes to the global cancer burden with regional variations in incidence and mortality [8]. These statistics underscore the critical need for accurate and scalable hematological diagnostics, as interpretation of blood films remains complex, prone to inter-observer variability [72], and limited by shortages of trained hematologists and resources, specifically in low- and middle-income countries [51, 83].

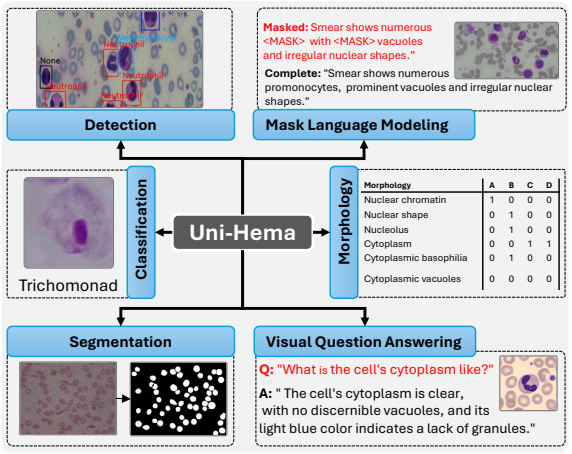

Figure 1. **Uni-Hema**: A unified architecture supporting diverse digital hematopathology tasks; including cell detection, morphology prediction, and cell segmentation in both single-cell and complete field-of-view (FoV) images, as well as visual question answering, and masked language modeling.

Recent advances in medical image analysis have transformed traditional blood-based microscopy examination into the emerging field of digital hematopathology [22]. Previously, most of the existing methods [2, 13, 33, 59, 70] were designed for single-task and single-disease applications, requiring separate datasets for each diagnostic purpose, thereby restricting scalability. Blood smear images further enhance this challenge due to overlapping cells, wide morphological diversity, and the need for multi-task

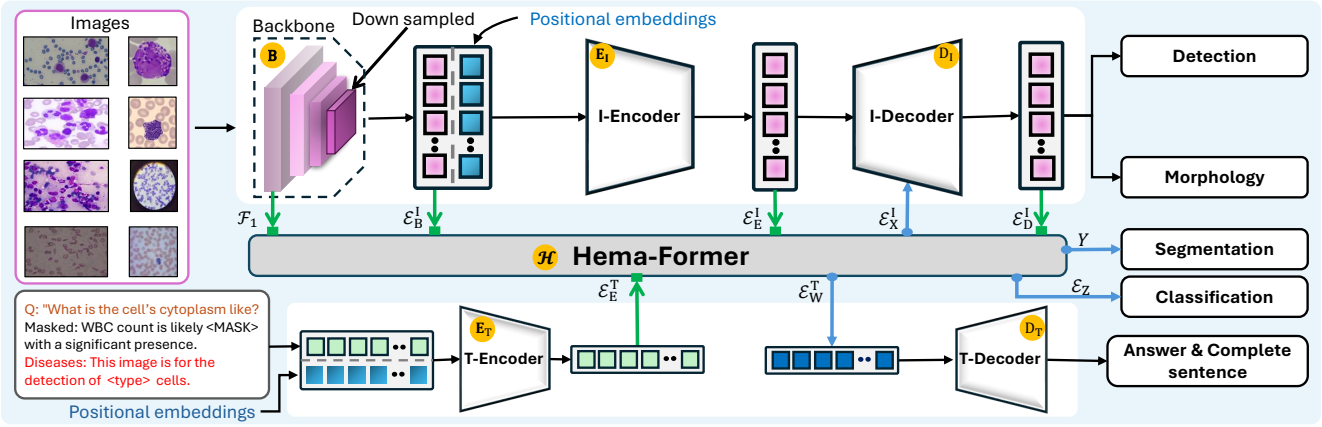

Figure 2. Uni-Hema model architecture comprises six principal modules: ($\mathbf{B}$) an image backbone for extracting spatial features, ($\mathbf{E_I}$) an image encoder for hierarchical visual embeddings, ($\mathbf{D_I}$) an image decoder for cell detection and morphology prediction; ($\mathbf{E_T}$) a text encoder for extracting textual features with respect to tasks, ($\mathbf{D_T}$) a text decoder for answer and sentence generation, ($\mathcal{H}$) and a Hema-former module, which serves as the core of the model by bridging visual and textual representations employing four submodules (see section 3).

processing, including detection, segmentation, classification, morphological interpretation, and visual–textual reasoning [64]. Furthermore, such task-specific approaches may restrict clinical applicability, as they struggle to adapt to the diverse and complex demands of real-world healthcare environments.[44, 58]. To overcome this issue, Moor *et.al* [44] have proposed Generalist Medical Artificial Intelligence (GMAI), which utilizes the recently developed medical foundation models [82] for the different medical tasks. Furthermore, the models in GMAI depend on natural language-based supervision [23, 45, 77]. Despite their success in vision-language tasks, these foundation models are limited in addressing essential vision-centric problems like detection and segmentation [10, 89]. The foundation models in digital pathology are primarily optimized for whole slide images (WSIs) of solid tissues captured at lower magnifications (5×, 40×) [9, 47, 79, 80]. In contrast, hematological diagnosis requires fine-grained, cell-level information at much higher magnifications (40×, 100×) [2, 5, 21, 31, 60]. Some recently introduced hematology foundation models [31, 87] are largely limited to single-cell tasks, thus unable to handle the realistic complexity of multiple cells (which might be overlapping or connecting), and even then are restricted by the cell type and tasks.

These limitations highlight a critical gap and emphasize the need for a *unified* multi-task, multi-modal (image and text), multi-disease model tailored specifically to the demands of digital hematopathology (DHP). Such a model should unify visual and textual understanding for digital hematopathology by integrating preferably existing datasets and performing diverse tasks (classification, detection, segmentation, and morphological reasoning), enabling comprehensive, cell-level interpretation across multiple hematological conditions such as leukemia, malaria, and anemia.

Developing unified models is challenging because it depends on having a complete benchmark for all the targeted tasks [28]. Most publicly available hematopathology datasets are disease-specific and task-limited, for example, focusing solely on malaria parasite detection [13, 18, 70] or leukocyte classification [2, 17, 33], thus lacking the diversity needed to generalize in multiple hematological disorders. In parallel, even in the pathology vision–language foundation models [9, 39, 47, 81] rely on paired image and caption datasets, which are difficult to create in hematopathology because they require precise, cell-level and clinically accurate descriptions [76, 79]. The complexity increases further in field-of-view images of a microscope due to the wide variation in the number of cells, their types, and morphological patterns [60].

To address these limitations, we propose *Uni-Hema*, a unified model designed for multi-task, multi-modal, and multi-disease settings of DHP. Uni-Hema has been designed to perform detection, classification, segmentation, morphology prediction, and cell-level visual question answering by leveraging the existing datasets. To handle multiple tasks, image types, a unified model should learn and capture representations across multiple hierarchical levels, while effectively conditioning on features from the image and text input. Uni-Hema embodies an image feature processing pipeline, consisting of a CNN and transformer-based encoding and decoding, and a text-based encoder and decoder network. Information across these seemingly parallel networks is carried out by an information mixture module called *Hema-Former*. Once features are extracted, these are forwarded to different task heads, ranging from cell detection (localization) and cell classification, morphological

feature prediction, binary image segmentation head, image classification, and answer & sentence completion head. Our contributions are summarized below.

- We propose **Uni-Hema**, the first unified vision-language model for digital hematopathology that jointly performs multiple tasks such as detection, segmentation, classification, and morphology reasoning across the six diverse disorders and diseases, including leukemia, malaria, anemia, and sickle cell disease.
- We have designed Hema-Former, a multi-modal feature fusion mechanism. It inputs feature representations of text and image at different hierarchical levels, and using an attention mechanism, generates features of different granularity appropriate for different tasks. These different feature generation strategies are implemented through learnable 4 sub-modules. (Section 3.2.2)
- We combine 46 diverse hematology datasets comprising 11 segmentation ($\approx 222K$), 17 detection ($\approx 84K$), and 16 classification ($\approx 380K$) datasets, along with curated 22K Question Answer pairs and 7K masked language modeling samples, establishing the most extensive multimodal corpus for digital Hematopathology to date.
- The unified model allows to learn representations that benefit from detection, segmentation, and single-cell interpretation tasks. Therefore, even when trained once jointly for all the tasks, the results are better or comparable to the single-task, single-dataset SOTA methods. This trend is visible on unseen datasets as well.

2. Related Work

The early approaches in digital hematopathology are limited to addressing single-task objectives, mostly focusing on single cell classification [2, 11, 16, 41], detection [13, 55, 70, 78], or segmentation [12, 32, 38, 43, 65, 90] independently. More recent research has extended toward dual-task and multi-task learning frameworks [52, 53, 59], enabling the joint optimization of related tasks such as classification and segmentation to enhance diagnostic precision and computational efficiency. Driven by rapid advances in deep learning, research increasingly focuses on developing VLMs [15, 27, 39], foundation models [31, 47, 81, 87], and unified frameworks [37] that generalize across tasks and modalities, moving beyond task-specific designs.

Vision Language Models (VLM's): The emergence of VLMs has bridged visual and textual modalities, enabling joint reasoning across computer vision and natural language processing tasks. Models such as CLIP [56], ALIGN [27], and CoCa [86] have pioneered task-agnostic pretraining by leveraging large-scale image–caption pair datasets to learn shared representation spaces, achieving strong generalization across diverse downstream tasks. Despite these advances, the lack of large-scale, high-quality image–text datasets specific to the sub-domains of pathology continues to constrain their generalization and downstream applicability. Pathology-oriented frameworks like CONCH [39], Transpath [79], and Pathology foundation model (PFM) [47] introduce domain adaptation strategies to improve multi-modal alignment for tissue slides. However, their transferability to hematology remains limited due to hematology-specific data scarcity, cell morphological diversity, and the fine-grained cellular variations that are different from pathology imaging modalities.

Foundation model: In the biomedical imaging domain, foundation models represent an effort to move beyond narrow task formulations by establishing transferable visual representations that can support a range of downstream analyses from tissue segmentation to cellular morphology classification. Approaches such as MedSAM [40], DinoBloom [31], and RedDino [87] demonstrate the growing potential of large-scale pretraining in medical and hematology contexts. However, these models remain limited to single-cell analysis and textual integration, lacking unified multimodal reasoning across detection and contextual reasoning. While multimodal foundation models (e.g., UNI [9], PFM [47]) emphasize image–text retrieval but lack fine-grained biomedical understanding.

Unified model: Recent advancements have accelerated the development of unified models capable of performing multiple tasks within a single framework. Models such as DINO-X [62], Uni-3DL [37], and Uni-Perceiver v2 [36] exemplify this trend by incorporating multimodal inputs and modality-agnostic transformers to enable scalable, task-agnostic learning across diverse domains. However, in the medical imaging domain, unified architectures remain limited. Existing approaches, such as CelloType [53] and LeukemiaAttri [59] demonstrate progress toward multi-task learning by jointly performing dual tasks, but still rely on task-specific single datasets. Nonetheless, these methods remain constrained to specific diseases with annotated multitask datasets and lack the broader adaptability required for a unified model for multi-task, multi-disease, multimodal hematological analysis. To address the above-mentioned limitation of VLMs, foundation and unified models for hematopathology, our proposed unified framework leverages diverse datasets to enable integrated learning against multiple tasks and hematology diseases.

3. Methodology

3.1. Dataset Formation

To train the unified model, we utilized 46 publicly available hematology datasets encompassing both single-cell and full-field microscopic images, covering a wide range of disease categories, including malaria, leukemia, anemia, sickle cell disease, and healthy samples. For segmentation or detection tasks, we generate a textual prompt for each im-

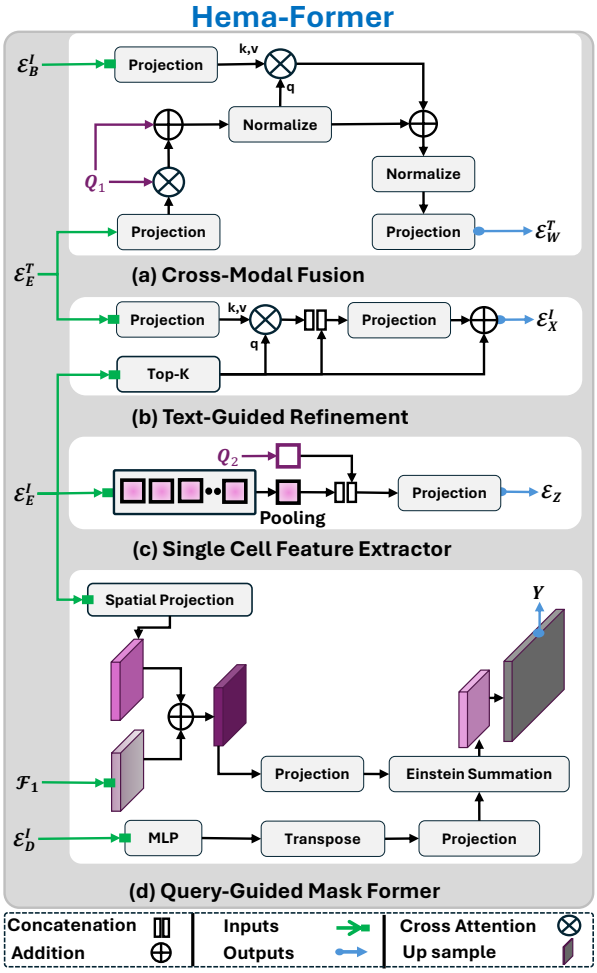

Figure 3. Hema-Former sub-modules: (a) Cross-modal fusion aligns text feature queries ($\mathcal{E}_{\mathbf{E}}^{\mathbf{T}}$) and image feature queries ($\mathcal{E}_{\mathbf{E}}^{\mathbf{I}}$), incorporating learnable queries ($\mathbf{Q}_1$) (b) Text-guided refinement that updates top-K object queries through cross-attention with text queries (c) Single cell feature extractor aggregates mean-pooled encoder features ($\mathcal{E}_{\mathbf{E}}^{\mathbf{I}}$) with learnable query ($\mathbf{Q}_2$) for single-cell classification, and (d) Query-guided mask former (QMF) module that utilizes the fused backbone features ($\mathcal{F}_1$) and spatially projected encoder features ($\mathcal{E}_{\mathbf{E}}^{\mathbf{I}}$) to predict segmentation masks ($\mathbf{Y}$).

age in the form: "This image is for the detection of <type of disease > of cells." For Question Answer, the prompt starts with **Q:**, and for Masked Language Modeling, the prompt starts with **mask:** tokens. The single-cell classification relied solely on visual features. The details of the datasets are given in Section 4.1.

3.2. Uni-Hema

The architecture of Uni-Hema ($\mathcal{U}$) is illustrated in Figure 2. It serves as an integrated vision–textual framework designed to address multiple tasks in digital hematopathology. The architecture comprises six principal modules: the backbone ($\mathbf{B}$), image encoder ($\mathbf{E_I}$), text encoder ($\mathbf{E_T}$), image decoder ($\mathbf{D_I}$), text decoder ($\mathbf{D_T}$), and Hema-Former module ($\mathcal{H}$). We first outline the notation and basic feature extraction steps used in UNI-Hema.

Given an input hematological image $\mathbf{I}$ and a text prompt $\mathbf{T}$, the image backbone $\mathbf{B}$ extracts multi-scale spatial features $\{\mathcal{F}_i\}_{i=1}^{L}$, where each $\mathcal{F}_i \in \mathbb{R}^{c_i \times m_i \times n_i}$ represents the i^{th} feature map with c_i channels and spatial resolution $m_i \times n_i$ and 'L' represents the number of levels. These features are converted into a spatial embedding space $\mathcal{E}_{\mathbf{B}}^{\mathbf{I}} \in \mathbb{R}^{V_t \times M}$ and encoded by a pre-trained encoder $\mathbf{E_I}$ to produce contextualized visual representations $\mathcal{E}_{\mathbf{E}}^{\mathbf{I}} \in \mathbb{R}^{V_t \times M}$, where V_t denotes the number of spatial tokens and M represents the embedding dimension. In parallel, text encoder $\mathbf{E_T}$ derives contextual embeddings $\mathcal{E}_{\mathbf{E}}^{\mathbf{T}} \in \mathbb{R}^{L_t \times N}$ from the text prompt $\mathbf{T}$, where L_t denotes the number of textual tokens and N is the text embedding dimension. Finally, the image decoder $\mathbf{D_I}$ refines the $\mathcal{E}_{\mathbf{E}}^{\mathbf{I}}$ features into disease-informed object embeddings $\mathcal{E}_{\mathbf{D}}^{\mathbf{I}}$ for downstream reasoning.

3.2.1. Image and Text Encoders:

Image Backbone: We employ a CNN-based ResNet [20] backbone ($\mathbf{B}$) as the visual feature extractor. The $\mathbf{B}$ generates multi-level spatial feature $\{\mathcal{F}_i\}_{i=1}^{L}$, capturing both fine-grained low-level information and high-level contextual details. These features are utilized for generating the segmentation masks, while projected spatial embeddings $\mathcal{E}_{\mathbf{B}}^{\mathbf{I}}$ serve as foundational representations for the other tasks.

Image Encoder: To enhance contextual understanding and capture long-range dependencies between spatial embeddings, inspired by DINO [88], a six-layer spatial image encoder $\mathbf{E_I}$ is integrated on top of backbone features. The $\mathbf{E_I}$ refines $\mathcal{E}_{\mathbf{B}}^{\mathbf{I}}$ into contextualized embeddings $\mathcal{E}_{\mathbf{E}}^{\mathbf{I}}$ using self-attention, enabling the model to encode both global and local structural relationships. Global features support robust classification of single cells and morphology, while local features capture objectness in the microscopy field of view.

Text Encoder: Textual branch of the $\mathcal{U}$ framework employs a transformer-based $\mathbf{E_T}$ derived from T5 [57]. $\mathbf{E_T}$ processes the text prompt $\mathbf{T}$ of hematological disease type, morphology-related questions, and masked sentences to produce contextualized textual embeddings $\mathcal{E}_{\mathbf{E}}^{\mathbf{T}}$. These embeddings capture underlying semantic relationships, enabling effective multimodal alignment and task-aware interaction with corresponding visual representations.

3.2.2. Hema-Former:

The proposed **Hema-Former** ($\mathcal{H}$) integrates inputs from both text and image encoders to produce hierarchical, task-specific fused embeddings. For visual question answering and masked language modeling, it aligns textual and visual representations to enable multimodal reasoning. For detection and segmentation tasks, it focuses on token-level attention to enhance targeted object localization, while for pixel-

level operations, it combines textual embeddings with fine-grained backbone features and image embeddings to generate accurate segmentation masks. Furthermore, a classification feature extractor enables text-independent image-level representation learning for classification. Figure 3 illustrates the architecture of the Hema-Former module, where the proposed Hema-Former bridges visual and textual representations through a set of integrated sub-modules.

a) Cross-modal fusion (CMF): For effective alignment between textual and visual semantic embeddings, the proposed CMF computes cross-attention correlations between $\mathcal{E}_{\mathbf{E}}^{\mathbf{T}}$, $\mathcal{E}_{\mathbf{B}}^{\mathbf{I}}$, facilitating unified understanding. Let the textual embeddings be $\mathcal{E}_{\mathbf{E}}^{\mathbf{T}} \in \mathbb{R}^{L_T \times N}$ and the visual features be $\mathcal{E}_{\mathbf{B}}^{\mathbf{I}} \in \mathbb{R}^{L_V \times M}$, learnable queries be about $\mathbf{Q}_1 \in \mathbb{R}^{L_V \times M}$. First, cross-attention is applied between $\mathbf{Q}_1$ and the projected textual embeddings $\mathcal{E}_{\mathbf{E}}^{\mathbf{T}\prime\prime} = \mathrm{P}(\mathcal{E}_{\mathbf{E}}^{\mathbf{T}})$:

$$\mathbf{J} = \mathrm{Norm}(\mathbf{Q}_2 + \mathrm{CrossAttn}(\mathbf{Q}_2, \mathcal{E}_{\mathbf{E}}^{\mathbf{T}\prime\prime}, \mathcal{E}_{\mathbf{E}}^{\mathbf{T}\prime\prime}), \quad (1)$$

where Norm represents the normalization layer applied to stabilize the distribution of features. Next, cross attention is applied between the $\mathbf{J}$ and projected visual features $\mathrm{P}(\mathcal{E}_{\mathbf{B}}^{\mathbf{I}})$, then fused with $\mathbf{J}$ and normalize as:

$$\mathcal{E}_{\mathbf{W}}^{\mathbf{T}} = \mathrm{Norm}(\mathbf{J} + \mathrm{CrossAttn}(\mathbf{J}, \mathrm{P}(\mathcal{E}_{\mathbf{B}}^{\mathbf{I}}), \mathrm{P}(\mathcal{E}_{\mathbf{B}}^{\mathbf{I}})), \quad (2)$$

where $\mathcal{E}_{\mathbf{W}}^{\mathbf{T}} \in \mathbb{R}^{L_t \times N}$ enriches the semantically aligned multimodal embeddings, enabling joint reasoning over visual and textual domains for the text decoder ($\mathbf{D}_T$) to generate language-driven tasks such as masked sentence completion and question answering.

b) Text-guided visual refinement (TGVR): To enable the distinction between different types of hematological disease images for the segmentation and detection tasks, the Top-K ($\mathbf{k}$) queries are separated from the $\mathcal{E}_{\mathbf{E}}^{\mathbf{I}}$ based on the objectness, similar to [88]. For the disease-attend visual representation, the textual embeddings $\mathcal{E}_{\mathbf{E}}^{\mathbf{T}}$ are processed jointly with the $\mathbf{k}$ separated visual queries. To align their dimensions, a projection layer is first applied as $\mathcal{E}_{\mathbf{E}}^{\mathbf{T}\prime} = \mathrm{P}(\mathcal{E}_{\mathbf{E}}^{\mathbf{T}})$, where $\mathrm{P}(\cdot)$ denotes the projection function. The resulting $\mathcal{E}_{\mathbf{E}}^{\mathbf{T}\prime}$ is then utilized as the key and value in a cross-attention operation, with $\mathbf{k}$ serving as the query. The attended features obtained from this operation are concatenated with $\mathbf{k}$ and passed through an additional projection layer to generate the refined decoder queries $\mathcal{E}_{\mathbf{X}}^{\mathbf{I}}$:

$$\mathcal{E}_{\mathbf{X}}^{\mathbf{I}} = \mathrm{P}\left([\mathbf{k} \,||\, \mathrm{CrossAttn}(\mathbf{k}, \mathcal{E}_{\mathbf{E}}^{\mathbf{T}\prime}, \mathcal{E}_{\mathbf{E}}^{\mathbf{T}\prime})]\right), \quad (3)$$

where $||$ indicates concatenation. $\mathcal{E}_{\mathbf{X}}^{\mathbf{I}} \in \mathbb{R}^{D_t \times M}$ represent the disease guided top-K queries for the image decoder ($\mathbf{D}_I$), to producing disease-informed object embeddings $\mathcal{E}_{\mathbf{D}}^{\mathbf{I}}$ that enable detection and morphology estimation and assist segmentation. D_t represent the dimension of the topk-K queries,

c) Single cell feature extractor (SCFE): To extract the global image-level features (independent of the text input), we design the SCFE module. In this module, a learned query is concatenated with the mean of the image embeddings and then passed through a projection layer to produce the image-level feature, denoted as $\mathcal{E}_{\mathbf{Z}} \in \mathbb{R}^{1 \times M}$.

d) Query-guided mask former (QGMF): This module generates the binary segmentation mask by integrating spatial features from the backbone and contextual embeddings from the image encoder. It further incorporates disease-informed object embeddings from the image decoder, following the design principles of Mask DINO [35].

Specifically, for the input image $\mathbf{I}$, spatial features $\mathcal{F}_1$ extracted from backbone are fused with the image embedding $\mathcal{E}_{\mathbf{E}}^{\mathbf{I}}$, extracted from image-encoder. In our current implementation, fusion is done by a summation operation, but could be replaced by any other operation. The fused features are then passed through a learnable projection layer to obtain $\mathbf{G}_{\text{proj}}$. In parallel, the disease-informed object query embeddings $\mathcal{E}_{\mathbf{D}}^{\mathbf{I}}$ are processed through a multilayer perceptron (MLP), followed by transposition and projection via $\mathbf{P}$:

$$\mathcal{E}_{\mathbf{D}}^{\mathbf{I}\,\prime} = \mathbf{P}\left(\mathrm{MLP}(\mathcal{E}_{\mathbf{D}}^{\mathbf{I}})\right)^{\top}, \quad (4)$$

These projected embeddings are then combined with the spatial feature maps $\mathbf{G}_{\text{proj}}$ through a cross-modal interaction implemented using Einstein summation ($\mathcal{S}$), resulting in segmentation logits:

$$\mathbf{Y} = \mathcal{S}(\mathcal{E}_{\mathbf{D}}^{\mathbf{I}\,\prime}, \mathbf{G}_{\text{proj}}), \quad (5)$$

where $\mathbf{Y} \in \mathbb{R}^{\mathbf{D}_t \times m \times n}$. For binary segmentation, the mask associated with the first embedding is selected as the final binary segmentation prediction $\mathbf{Y} \in \mathbb{R}^{1 \times m \times n}$.

3.2.3. Image and Text Decoders:

Image decoder: The image decoder ($\mathbf{D_I}$) in Uni-Hema extends the DINO Detr [88] architecture. The output $\mathcal{E}_{\mathbf{D}}^{\mathbf{I}}$ of the image decoder $\mathbf{D_I}$ serves as a shared representation for the three tasks, i.e., segmentation (through QGMF) and supporting cell detection and morphology prediction tasks.

Text decoder: The text decoder in Uni-Hema builds upon the transformer-based architecture [57]. It generates textual outputs autoregressively, attending to both previously decoded tokens and the fused multimodal representations through cross-attention with the visual encoder features. This mechanism allows the model to perform domain-specific reasoning tasks such as masked sentence completion and single-cell-based question answering.

4. Experiments

4.1. Datasets

We curated a large collection of publicly available hematology datasets, preserving existing annotations for consis-

tency. In total, 18 datasets were used for detection, 11 for segmentation, and 17 for classification, totaling ≈ 0.7 million images across diverse diseases. Two morphologically annotated datasets [59, 76] are used to curate the visual Question Answering and masked language modeling tasks.

Segmentation : The cell segmentation datasets encompass malaria-infected cell datasets such as Malaria Cell Images [1], ErythrocytesIDB [19], and MP-IDB [38]; white blood cell segmentation datasets including BBBC041Seg [12], NuClick [32], KRD-WBC [3], WBC Image dataset [90], and the White Blood Cell dataset [43]; as well as anemia and red blood cell datasets [14, 65].

Single Cell for classification: The single-cell (SC) classification datasets, comprising both white blood cells (WBCs) and red blood cells (RBCs), were collected from a diverse range of publicly available studies and repositories. These include peripheral blood smear datasets [2–4, 7, 21, 32, 33, 41, 46, 52, 67, 68] and bone marrow datasets [42]. In total, approximately 365k high-resolution single-cell images were utilized, spanning nearly 45 morphologically distinct WBC and RBC classes.

Cell Detection: Our collection includes several large-scale and diverse hematological microscopy full microscopic field of view datasets for the cell detection, such as the largest leukemia dataset [59] that includes morphology attributes, along with M5 [70], TXL-PBC [17], BCCD [68], Sickle-cell [78], Erythrocytes [19], MP-IDB [38], PD-PF [74], Plasmodium [55], Vivax [24], Raabin [33], AneRBC [65], Bio-Net [66], ThickBloodSmears [84], and NIH-NLM-Thick PV [30]. In total, this collection comprises approximately 84k images representing 30 distinct cell classes, covering a wide range of disease categories.

MLM and VQA: To generate descriptive sentences from microscopy data, we employed medical language model, BioMistral 7B [34] on the WBCAtt dataset [76] to construct question–answer pairs to single-cell reasoning and multimodal understanding. Similarly, Gemini 1.5 [73] was applied to full field-of-view images from the LeukemiaAttri dataset [59], leveraging morphological annotations to create both masked and fully descriptive captions. We named both datasets as WBCAtt-VQA and LeukemiaAttri-MLM, respectively. The generation process for both datasets is semi-synthetic, where detailed ground truth morphological annotations are utilized as prompts to guide descriptive synthesis. The generated QA and masked sentence are verified through context-based evaluation to ensure semantic coherence and linguistic accuracy. Note that this dataset is curated only for experimental purposes and is not recommended for medical use.

4.2. Implementation details

In this work, we use ResNet-50 [20] as the backbone, and adopt the DINO-DETR [88] architecture with six transformer layers in both the encoder and decoder, initialized with pretrained weights. For textual processing, we employ encoder-decoder based T5 [57] initialized with `t5_base` weights. Training is performed in six sequential stages.

Step 1: Pre-train the image backbone and encoder on single-cell (SC) classification datasets for 24 epochs with a batch size of 32. **Step 2:** Pre-train the text encoder and decoder on semi-synthetic masked language modeling (MLM) and question-answering (QA) datasets for the medical context learning. **Step 3:** Jointly train the vision modules of the Uni-Hema on classification, segmentation, and detection datasets using two images per task per batch (total batch size = 6) for 12 epochs, updating the visual modules, including the TGR, SCFE, and QGMF modules, simultaneously. The rest of the modules ($\mathbf{E_T}$, $\mathbf{D_T}$, CMF) are frozen. **Step 4:** Fine-tune the image decoder for 12 epochs on all detection and microscopy-field-of-view (FoV) morphology datasets. **Step 5:** Fine-tune the query-guided mask former (QGMF) on the complete segmentation dataset for 8 epochs. **Step 6:** Train the cross-modal fusion (CMF) and $\mathbf{D_T}$ on combined VQA and V-MLM datasets for 24 epochs, all other modules and submodules freeze. The complete training process requires approximately 8 days on a single NVIDIA RTX 4090 GPU (the best available resource that we have). The details of the training configuration for each step are provided in the Supplementary material.

4.3. Results and Discussion

For detection, we report comparisons with DINO [88] and YOLO [29]. For segmentation, we include U-Net, NanoNet [25], and TransNetR [26]. In single-cell classification, we compare against DINOv2S [49], DINOBloomS [31], and DINOv3S [69]. For multi-task comparison, the performance of AttriDet [59] on FoV morphology and cell detection is reported on the LeukemiaAttri dataset. To evaluate VQA and MLM task, in-house curated datasets WBCAtt-VQA and LeukemiaAttri-MLM are used. Other results and comparisons for detection (Faster R-CNN [61], Sparse R-CNN [71], FCOS [75]), segmentation (U-Net, Attention U-Net [48], and classification (TransPath [79], CONCH [39], Phikon-v2 [15], ResNet-50 [20], UNI [9]) are all reported in the Supplementary Material.

Note that in the following experiments, each baseline method is trained on its corresponding dataset individually, whereas Uni-Hema is jointly trained on the entire training corpus.

Detection Results: We compare the performance of cell detection across multiple datasets with previous state-of-the-art fully supervised methods as shown in Table 1. The results demonstrate that our proposed unified method achieves superior or comparable performance on cell detection tasks across various diseases, including leukemia, malaria, sickle-cell, and normal individuals. For sickle cell

Table 1. Performance of Uni-Hema across multiple tasks. Baseline methods are trained separately for each task and for each dataset. Uni-Hema, on the other hand, is trained only once and tested on the tasks without any re-training or fine-tuning. A "–" symbol indicates that a baseline model does not support the corresponding task. Values in bold and underlined denote the best and second-best results.

Dataset	DINO	YOLO	U-Net	Nanonet	TransNetR	DinoBloomS	Dinov2S	Dinov3s	Resnet	AttriDet	Ours	*Disease*
Cell Detection							**mAP$_{50}$**					
H_40x_C2 [59]	36.9	<u>37.3</u>	-	-	-	-	-	-	-	40.6	**43.6**	Leukemia
H_100x_C2 [59]	43.7	<u>44.2</u>	-	-	-	-	-	-	-	44.4	**49.8**	Leukemia
L_40x_C2 [59]	<u>36.6</u>	34.9	-	-	-	-	-	-	-	35.4	**40.7**	Leukemia
L_100x_C2 [59]	<u>38.2</u>	38.1	-	-	-	-	-	-	-	36.2	**45.6**	Leukemia
H_1000x [70]	<u>79.8</u>	77.3	-	-	-	-	-	-	-	-	**83.1**	Malaria
L_1000x [70]	**64.2**	56.5	-	-	-	-	-	-	-	-	<u>62.4</u>	Malaria
H_400x [70]	**70.4**	66.9	-	-	-	-	-	-	-	-	<u>69.0</u>	Malaria
L_400x [70]	58.3	**59.9**	-	-	-	-	-	-	-	-	<u>54.5</u>	Malaria
Sickle Cell [78]	**73.6**	<u>68.6</u>	-	-	-	-	-	-	-	-	67.0	Sickle Cell
Parasites [13]	<u>38.6</u>	**46.1**	-	-	-	-	-	-	-	-	36.2	Parasites
BCCD [68]	**89.5**	<u>88.1</u>	-	-	-	-	-	-	-	-	87.8	Normal
TXL [17]	**95.3**	<u>94.9</u>	-	-	-	-	-	-	-	-	94.0	Normal
Mean	60.4	59.4	-	-	-	-	-	-	-	-	39.1	**61.1**
Single Cell Classification							**F1**					
Raabin [33]	-	-	-	-	-	<u>98.0</u>	93.7	93.8	88.1	-	**98.8**	Normal
BMC[42]	-	-	-	-	-	<u>85.0</u>	68.2	67.8	64.7	-	**86.2**	Lymphoma
Mean	-	-	-	-	-	91.5	81.0	90.8	76.4	-	**92.5**	-
Segmentation							**Dice-Score**					
AneRBC-Anemic[65]	-	-	78.3	91.2	**93.6**	-	-	-	-	-	<u>93.4</u>	Anemia
AneRBC-Healthy[65]	-	-	75.1	90.9	**95.2**	-	-	-	-	-	<u>94.1</u>	Normal
Elsafty [14]	-	-	93.4	98.3	<u>99.5</u>	-	-	-	-	-	**99.9**	Anemia
IDB2[38]	-	-	91.5	33.1	**92.1**	-	-	-	-	-	<u>90.5</u>	Leukemia
KRD[3]	-	-	93.3	86.7	**94.9**	-	-	-	-	-	<u>94.5</u>	Unknown
MD-2019[1]	-	-	75.2	84.7	**86.7**	-	-	-	-	-	77.6	Malaria
Mean	-	-	84.5	80.8	**93.7**	-	-	-	-	-	<u>91.7</u>	-
Cell Morphology (FoV)							**F1**					
H_40x_C2 [59]	-	-	-	-	-	-	-	-	-	64.1	**75.0**	Leukemia
H_100x_C2 [59]	-	-	-	-	-	-	-	-	-	71.1	**75.8**	Leukemia
L_40x_C2 [59]	-	-	-	-	-	-	-	-	-	53.2	**74.5**	Leukemia
L_100x_C2 [59]	-	-	-	-	-	-	-	-	-	61.8	**83.6**	Leukemia
Mean	-	-	-	-	-	-	-	-	-	62.6	**77.2**	-
Single Cell Visual QA							**BLEU-4**					
WBCAtt-VQA	-	-	-	-	-	-	-	-	-	-	**56.4**	Normal
Masked Language Modeling							**BLEU-4**					
LeukemiaAttri-MLM	-	-	-	-	-	-	-	-	-	-	**79.8**	Leukemia

and parasite detection datasets, the number of training samples is very limited, making it difficult to learn dataset-specific patterns within a large corpus. In contrast, when training on individual datasets, the reduced complexity allows the model to learn more effectively.

Single-Cell Classification Results: For a fair comparison, we use Uni-Hema as a feature extractor, obtaining $\mathcal{E}_{\mathbf{X}}$ features from its SCFE module, and evaluate them against pre-trained state-of-the-art (SOTA) feature extractors. These include non-medical-domain models [20, 49, 69], medical-domain models [9, 15, 39, 79], and a hematology-specific vision foundation model designed for white blood cell images, DinoBloom [31]. Experimental results Table 1 demonstrate that Uni-Hema has superior performance compared to existing SOTA methods on both 21-class BMC [42], and 5-class Raabin [33] dataset.

Segmentation Results: To comprehensively assess segmentation performance, we compare Uni-Hema with leading fully supervised methods trained and tested on the particular dataset, including U-Net [63], TransNetR [26], and NanoNet [25], as shown in Table 1. Our approach achieves competitive results, as reflected by the *Dice score*. Notably, the reported performance corresponds to a single unified model evaluated across all datasets, without any task-specific retraining. Instead of a SOTA decoder for upsampling, we rely on simple mask upscaling during training (step 3) and a small upsampler during fine-tuning (step 5) due to resource constraints. Although this may slightly affect edge sharpness, our results remain comparable to the state-of-the-art TransNetR method. A qualitative comparison with TransNetR [26] is illustrated in Figure 4, while additional comparisons with other methods [25, 48, 63] are provided in the Supplementary Material.

Morphology Results: Cellular morphology across the microscopy field of view (FoV) is evaluated using the dataset from [59] which contains 6 morphological attributes. As shown in Table 1, our method demonstrates superior performance compared to the corresponding SOTA multi-task, AttriDet [59] method.

To broaden the capability of our proposed unified model,

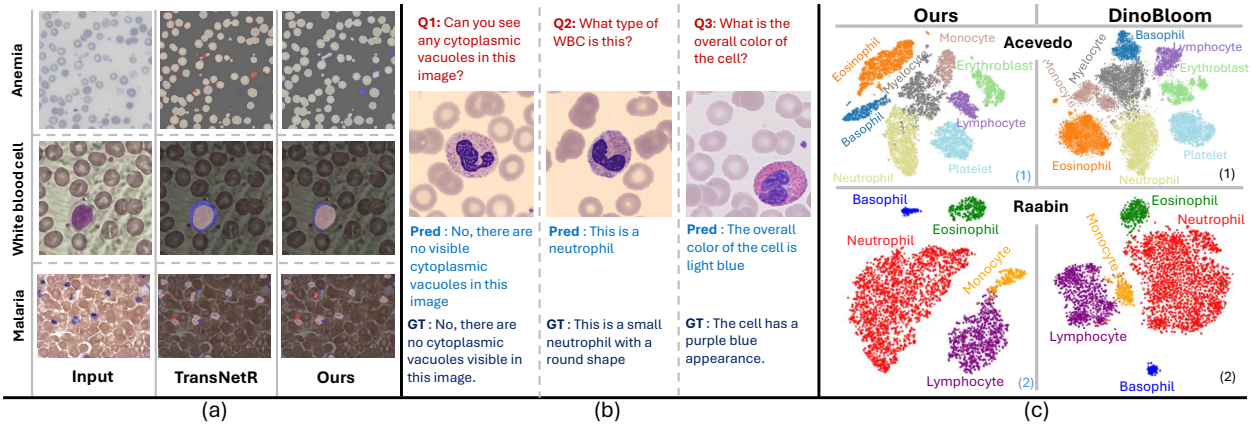

Figure 4. (a) Segmentation results of TransNetR and our method on anemia, malaria, and WBC images. TP, FP, and FN are shown in light yellow, blue, and red. Our method reduces false detections and improves localization, especially for anemia and WBC, while handling malaria robustly. (b) VQA outputs show contextually accurate predictions aligned with ground truth. (c) t-SNE plots of DinoBloom-S and Uni-Hema features on the Acevedo (8-class) and Raabin (5-class) datasets display clearer class separation with Uni-Hema.

Table 2. Comparison on unseen datasets: Uni-Hema demonstrates cross-domain adaptability across multiple tasks.

Dataset	DinoBloomS	Dinov2S	Dinov3S	ResNet	Ours
Cell Classification (F1)					
Acevedo [2]	**98.2**	94.5	96.0	90.0	98.1
MiniSit [85]	**98.8**	96.8	97.1	95.2	98.6
C-NMC [46]	69.6	71.0	68.0	64.9	**72.8**
RV-PBS_8 [52]	92.8	90.7	90.8	86.1	**93.6**
Mean	89.9	88.2	88.0	84.1	**90.8**
Detection (mAP_{50})					
Bio-Net [66]	-	-	-	-	54.7
Malaria [38, 66]	-	-	-	-	78.5
Cell Segmentation (Dice)					
BBBC041Seg [12, 76]	-	-	-	-	86.2
Cell Morphology (Single) — F1-Macro					
WBCAtt [76]	87.6	87.8	91.4	90.2	**91.6**

we incorporate two text-aligned visual understanding tasks, single-cell visual Question Answering (VQA) and full field-of-view Visual Masked Language Modeling (VMLM), into the training framework. As shown in Table 1, our UNI-Hema demonstrates strong performance across both tasks, achieving BLEU-4 scores of 56.4 for VQA, and 79.8 for V-MLM, respectively. These results highlight the model's ability to bridge textual reasoning, including the visual interpretation in hematological contexts. Qualitative examples in Figure 4 further illustrate the model's predictions.

Results on unseen datasets: We evaluate Uni-Hema on unseen datasets, and the results are shown in Table 2. For the classification task, we follow DinoBloom [31] experimental setup. Uni-Hema demonstrates overall superior performance with a mean F1-score of 90.8. For the cell detection task, it achieves a reasonable mAP@50 of 54.7 on the Bio-Net dataset [66] and 78.5 on the Malaria dataset [38]. Furthermore, for segmentation on unseen datasets, Uni-Hema attains a Dice score of 86.2, highlighting its strong robustness across multiple tasks. For single-cell morphological analysis, we follow the WBCAtt pipeline [76], treating the dataset as unseen for all feature extractor models [20, 31, 49, 69] listed in Table 2. The results show that Uni-Hema achieves state-of-the-art performance on single-cell morphology prediction. Detailed experimental results, including statistics, are given in the Supplementary Material.

Ablation studies: In the single-cell classification task, we integrate features from both the image backbone and the single-cell feature composer. On the unseen domain-shift dataset of Acevedo [2], this integration improves accuracy from 97.9% to 98.1%, and on the C-NMC dataset [46], from 68.8% to 72.8%. For the segmentation task, replacing simple interpolation-based upsampling with a small learnable upsampler network consistently enhances performance, improving segmentation results on the anemic dataset from 91.8% to 93.4% and on the healthy dataset from 92.8% to 94.1%, with similar gains observed across other datasets. Additional results are given in the supplementary material.

5. Conclusion

In this work, we present Uni-Hema, a unified model for multi-task, multi-disease, and multi-modal digital hematopathology. Uni-Hema integrates detection, segmentation, classification, morphological reasoning, and visual–textual understanding in a single architecture. A dedicated Hema-Former module aligns hierarchical visual features with task-driven textual features for contextual understanding. We further curate VQA and MLM datasets specialized for hematopathology to support the capabilities. Trained on 46 heterogeneous datasets with a progressive strategy, Uni-Hema matches or surpasses state-of-the-art single-task models.

Acknowledgments:

We express our sincere appreciation to the research team of the Intelligent Machine Lab, particularly Muhammad Ali and Nimra Dilawar, for their valuable contributions in reviewing and proofreading parts of this work. Furthermore, we acknowledge Information Technology University and Google for their partial funding support for this project. The code is publicly available at `https://github.com/intelligentMachines-ITU/Uni-Hema`.

References

[1] Syed Saiden Abbas and Tjeerd MH Dijkstra. Detection and stage classification of plasmodium falciparum from images of giemsa stained thin blood films using random forest classifiers. *Diagnostic pathology*, 15(1):130, 2020.

[2] 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.

[3] Haval Taha Ali, Fattah Alizadeh, and Nawsherwan Sadiq Mohammad. White blood cell microscopic image dataset for segmentation. *Available at SSRN 4617448*.

[4] JR Alipo-on, FI Escobar, JL Novia, MM Atienza, S Manaay, MJ Tan, N AlDahoul, and E Yu. Dataset for machine learning-based classification of white blood cells of the juvenile visayan warty pig. 2022.

[5] Md Zahangir Alom, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. Nuclei segmentation with recurrent residual convolutional neural networks based u-net (r2u-net). In *NAECON 2018-IEEE National Aerospace and Electronics Conference*, pages 228–233. IEEE, 2018.

[6] Barbara J Bain and Mike Leach. *Blood cells: a practical guide*. John Wiley & Sons, 2025.

[7] 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.

[8] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 74(3):229–263, 2024.

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

[10] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey E Hinton. A unified sequence interface for vision tasks. *Advances in Neural Information Processing Systems*, 35:31333–31346, 2022.

[11] Yifei Chen, Zhu Zhu, Shenghao Zhu, Linwei Qiu, Binfeng Zou, Fan Jia, Yunpeng Zhu, Chenyan Zhang, Zhaojie Fang, Feiwei Qin, et al. Sckansformer: Fine-grained classification of bone marrow cells via kansformer backbone and hierarchical attention mechanisms. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[12] Deponker Sarker Depto, Shazidur Rahman, Md Mekayel Hosen, Mst Shapna Akter, Tamanna Rahman Reme, Aimon Rahman, Hasib Zunair, M Sohel Rahman, and MRC Mahdy. Automatic segmentation of blood cells from microscopic slides: a comparative analysis. *Tissue and Cell*, 73: 101653, 2021.

[13] Eden. parasite detection dataset. `https://universe.roboflow.com/eden-lcx9y/parasite-detection`, 2024. visited on 2025-10-14.

[14] Ahmed Elsafty, Ahmed Soliman, and Yomna Ahmed. 1 million segmented red blood cells with 240 k classified in 9 shapes and 47 k patches of 25 manual blood smears. *Scientific Data*, 11(1):722, 2024.

[15] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024.

[16] Hüseyin Fırat. Classification of microscopic peripheral blood cell images using multibranch lightweight cnn-based model. *Neural Computing and Applications*, 36(4):1599–1620, 2024.

[17] Lu Gan and Xi Li. Txl-pbc: a freely accessible labeled peripheral blood cell dataset. *arXiv preprint arXiv:2407.13214*, 2024.

[18] GBD 2021 Sickle Cell Disease Collaborators. Global, regional, and national prevalence and mortality burden of sickle cell disease, 2000–2021: a systematic analysis from the global burden of disease study 2021. *The Lancet Haematology*, 10(8):e585–e599, 2023. Epub 2023 Jun 15. Erratum in: Lancet Haematol. 2023 Aug;10(8):e574. doi:10.1016/S2352-3026(23)00215-6.

[19] Manuel Gonzalez-Hidalgo, FA Guerrero-Pena, 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, 2014.

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

[21] Matthias Hehr, Ario Sadafi, Christian Matek, Peter Lienemann, Christian Pohlkamp, Torsten Haferlach, Karsten Spiekermann, and Carsten Marr. A morphological dataset of white blood cells from patients with four different genetic aml entities and non-malignant controls (aml-cytomorphology_mll_helmholtz). *(No Title)*, 2023.

[22] Yongfei Hu, Yinglun Luo, Guangjue Tang, Yan Huang, Juanjuan Kang, and Dong Wang. Artificial intelligence and its applications in digital hematopathology. *Blood Science*, 4 (3):136–142, 2022.

[23] 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.

[24] Jane Hung and Anne Carpenter. Applying faster r-cnn for object detection on malaria images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 56–61, 2017.

[25] Debesh Jha, Nikhil Kumar Tomar, Sharib Ali, Michael A Riegler, Håvard D Johansen, Dag Johansen, Thomas de Lange, and Pål Halvorsen. Nanonet: Real-time polyp segmentation in video capsule endoscopy and colonoscopy. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 37–43. IEEE, 2021.

[26] Debesh Jha, Nikhil Kumar Tomar, Vanshali Sharma, and Ulas Bagci. Transnetr: transformer-based residual network for polyp segmentation with multi-center out-of-distribution testing. In *Medical Imaging with Deep Learning*, pages 1372–1384. PMLR, 2024.

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[28] Sheng Jin, Shuhuai Li, Tong Li, Wentao Liu, Chen Qian, and Ping Luo. You only learn one query: learning unified human query for single-stage multi-person multi-task human-centric perception. In *European Conference on Computer Vision*, pages 126–146. Springer, 2024.

[29] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, yxNONG, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain. ultralytics/yolov5: v7.0 - yolov5 sota realtime instance segmentation, 2022.

[30] Yasmin M Kassim, Feng Yang, Hang Yu, Richard J Maude, and Stefan Jaeger. Diagnosing malaria patients with plasmodium falciparum and vivax using deep learning for thick smear images. *Diagnostics*, 11(11):1994, 2021.

[31] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 520–530. Springer, 2024.

[32] Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. Nuclick: a deep learning framework for interactive segmentation of microscopic images. *Medical Image Analysis*, 65:101771, 2020.

[33] Zahra Mousavi Kouzehkanan, Sepehr Saghari, Eslam Tavakoli, Peyman Rostami, Mohammadjavad Abaszadeh, Farzaneh Mirzadeh, Esmaeil Shahabi Satlsar, Maryam Gheidishahran, Fatemeh Gorgi, Saeed Mohammadi, et al. Raabin-wbc: a large free access dataset of white blood cells from normal peripheral blood. *bioRxiv*, pages 2021–05, 2021.

[34] Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

[35] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3041–3050, 2023.

[36] Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, et al. Uni-perceiver v2: A generalist model for large-scale vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2691–2700, 2023.

[37] Xiang Li, Jian Ding, Zhaoyang Chen, and Mohamed Elhoseiny. Uni3dl: Unified model for 3d and language understanding. *arXiv preprint arXiv:2312.03026*, 2023.

[38] Andrea Loddo, Cecilia Di Ruberto, Michel Kocher, and Guy Prod'Hom. Mp-idb: the malaria parasite image database for image processing and analysis. In *Sipaim–Miccai Biomedical Workshop*, pages 57–65. Springer, 2018.

[39] Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19764–19775, 2023.

[40] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024.

[41] Christian Matek, Simone Schwarz, Carsten Marr, and Karsten Spiekermann. A single-cell morphological dataset of leukocytes from aml patients and non-malignant controls (aml-cytomorphology_lmu). *The Cancer Imaging Archive (TCIA)[Internet]*, 2019.

[42] 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, The Journal of the American Society of Hematology*, 138(20):1917–1927, 2021.

[43] Mostafa Mohamed and Behrouz Far. An enhanced threshold based technique for white blood cells nuclei automatic segmentation. In *2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 202–207. IEEE, 2012.

[44] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[45] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[46] 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.

[47] Mieko Ochi, Daisuke Komura, and Shumpei Ishikawa. Pathology foundation models. *JMA journal*, 8(1):121–130, 2025.

[48] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[50] World Health Organization. *WHO Malaria Policy Advisory Group (MPAG) meeting report, 4, 5 and 7 March 2024*. World Health Organization, 2024.

[51] Debanjan Pain, Emily MacDuffie, Yehoda M Martei, Megan Kassick, Daniel J Ikeda, Lawrence N Shulman, Lina Loaiza Salazar, Dayssy Diaz Pardo, Shona Nag, and Surbhi Grover. Barriers to implementing a quality improvement program in low-and middle-income countries: adequacy of resources. *JCO Global Oncology*, 10:e2400114, 2024.

[52] 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.

[53] Minxing Pang, Tarun Kanti Roy, Xiaodong Wu, and Kai Tan. Cellotype: a unified model for segmentation and classification of tissue images. *Nature methods*, 22(2):348–357, 2025.

[54] William R Platt. Color atlas and textbook of hematology. *(No Title)*, 1969.

[55] John A Quinn, Alfred Andama, Ian Munabi, and Fred N Kiwanuka. Automated blood smear analysis for mobile malaria diagnosis. *Mobile point-of-care monitors and diagnostic device design*, page 115, 2018.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[57] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[58] Pranav Rajpurkar and Matthew P Lungren. The current and future state of ai interpretation of medical images. *New England Journal of Medicine*, 388(21):1981–1990, 2023.

[59] 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 553–563. Springer, 2024.

[60] Abdul Rehman, Talha Meraj, Aiman Mahmood Minhas, Ayisha Imran, Mohsen Ali, Waqas Sultani, and Mubarak Shah. Leveraging sparse annotations for leukemia diagnosis on the large leukemia dataset. *arXiv preprint arXiv:2504.02602*, 2025.

[61] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[62] Tianhe Ren, Yihao Chen, Qing Jiang, Zhaoyang Zeng, Yuda Xiong, Wenlong Liu, Zhengyu Ma, Junyi Shen, Yuan Gao, Xiaoke Jiang, et al. Dino-x: A unified vision model for open-world object detection and understanding. *arXiv preprint arXiv:2411.14347*, 2024.

[63] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[64] Muhammad Shahzad, Farman Ali, Syed Hamad Shirazi, Assad Rasheed, Awais Ahmad, Babar Shah, and Daehan Kwak. Blood cell image segmentation and classification: a systematic review. *PeerJ Computer Science*, 10:e1813, 2024.

[65] 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*, 2024:baae120, 2024.

[66] Usman Ali Shams, Isma Javed, Muhammad Fizan, Aqib Raza Shah, Ghulam Mustafa, Muhammad Zubair, Yehia Massoud, Muhammad Qasim Mehmood, and Muhammad Asif Naveed. Bio-net dataset: Ai-based diagnostic solutions using peripheral blood smear images. *Blood Cells, Molecules, and Diseases*, 105:102823, 2024.

[67] Eugene Shenderov. Acute promyelocytic leukemia (apl). Kaggle Dataset, 2019. Accessed: Aug. 6, 2025.

[68] shenggan, Nicolas Chen, cosmicad, and akshaylamba. Bccd: Blood cell count and detection, 2018.

[69] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[70] Waqas Sultani, Wajahat Nawaz, Syed Javed, Muhammad Sohail Danish, Asma Saadia, and Mohsen Ali. Towards low-cost and efficient malaria detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20655–20664. IEEE, 2022.

[71] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021.

[72] Rohollah Moosavi Tayebi, Youqing Mu, Taher Dehkharghanian, Catherine Ross, Monalisa Sur, Ronan Foley, Hamid R Tizhoosh, and Clinton JV Campbell. Automated bone marrow cytology using deep learning to generate a histogram of cell types. *Communications medicine*, 2(1):45, 2022.

[73] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[74] F Boray Tek, Andrew G Dempster, and Izzet Kale. Computer vision for microscopy diagnosis of malaria. *Malaria journal*, 8(1):153, 2009.

[75] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019.

[76] Satoshi Tsutsui, Winnie Pang, and Bihan Wen. Wbcatt: a white blood cell dataset annotated with detailed morphological attributes. *Advances in Neural Information Processing Systems*, 36:50796–50824, 2023.

[77] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

[78] Florence Tushabe, Samuel Mwesige, Vicent Kasule, Emily Nsiimire, Sarah C Musani, David Areu, and Emmanuel Othieno. An image-based sickle cell detection method. *Authorea Preprints*, 2024.

[79] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–195. Springer, 2021.

[80] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035): 970–978, 2024.

[81] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

[82] Walter F Wiggins and Ali S Tejani. On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence*, 4(4):e220119, 2022.

[83] Ying Xing, Xuekai Liu, Juhua Dai, Xiaoxing Ge, Qingchen Wang, Ziyu Hu, Zhicheng Wu, Xuehui Zeng, Dan Xu, and Chenxue Qu. Artificial intelligence of digital morphology analyzers improves the efficiency of manual leukocyte differentiation of peripheral blood. *BMC Medical Informatics and Decision Making*, 23(1):50, 2023.

[84] Feng Yang, Mahdieh Poostchi, Hang Yu, Zhou Zhou, Kamolrat Silamut, Jian Yu, Richard J Maude, Stefan Jaeger, and Sameer Antani. Deep learning for smartphone-based malaria parasite detection in thick blood smears. *IEEE journal of biomedical and health informatics*, 24(5):1427–1438, 2019.

[85] 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(1):41, 2023.

[86] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[87] Luca Zedda, Andrea Loddo, Cecilia Di Ruberto, and Carsten Marr. Reddino: A foundation model for red blood cell analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 445–455. Springer, 2025.

[88] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[89] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *Advances in Neural Information Processing Systems*, 35:36067–36080, 2022.

[90] 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.