Rank-Aware Agglomeration of Foundation Models for Immunohistochemistry Image Cell Counting

Zuqi Huang[a,b,1], Mengxin Tian[c,1], Huan Liu[d,*], Wentao Li[a], Baobao Liang[e], Jie Wu[f], Fang Yan[g], Zhaoqing Tang[c,h,*], Zhongyu Li[b,*]

[a]*School of Software Engineering, Xi'an Jiaotong University, Xi'an, 710049, Shaanxi, China*
[b]*School of Computer Science, Shanghai Jiao Tong University, Shanghai, 200240, China*
[c]*Department of Gastrointestinal Surgery and the Gastric Cancer Center, Zhongshan Hospital, Fudan University, Shanghai, 200032, China*
[d]*School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, Shaanxi, China*
[e]*The Comprehensive Breast Care Center, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710114, Shaanxi, China*
[f]*Department of Pathology, The Second Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710114, Shaanxi, China*
[g]*Shanghai Artificial Intelligence Laboratory, Shanghai, China*
[h]*Department of General Surgery, Zhongshan Hospital (Xiamen), Fudan University, Xiamen, 361015, Fujian, China*

Abstract

Accurate cell counting in immunohistochemistry (IHC) images is critical for quantifying protein expression and aiding cancer diagnosis. However, the task remains challenging due to the chromogen overlap, variable biomarker staining, and diverse cellular morphologies. Regression-based counting methods offer advantages over detection-based ones in handling overlapped cells, yet rarely support end-to-end multi-class counting. Moreover, the potential of foundation models remains largely underexplored in this paradigm. To address these limitations, we propose a rank-aware agglomeration framework that selectively distills knowledge from multiple strong foundation models, leveraging their complementary representations to handle IHC heterogeneity and obtain a compact yet effective student model, CountIHC. Unlike

*Corresponding authors: Zhongyu Li (zhongyuli@sjtu.edu.cn), Zhaoqing Tang (tang.zhaoqing@zs-hospital.sh.cn), Huan Liu (huanliu@xjtu.edu.cn).

[1]Equal contribution.

prior task-agnostic agglomeration strategies that either treat all teachers equally or rely on feature similarity, we design a Rank-Aware Teacher Selecting (RATS) strategy that models global-to-local patch rankings to assess each teacher's inherent counting capacity and enable sample-wise teacher selection. For multi-class cell counting, we introduce a fine-tuning stage that reformulates the task as vision–language alignment. Discrete semantic anchors derived from structured text prompts encode both category and quantity information, guiding the regression of class-specific density maps and improving counting for overlapping cells. Extensive experiments demonstrate that CountIHC surpasses state-of-the-art methods across 12 IHC biomarkers and 5 tissue types, while exhibiting high agreement with pathologists' assessments. Its effectiveness on H&E-stained data further confirms the scalability of the proposed method.

Keywords:
Rank-aware agglomeration, Foundation model, Cell counting, Immunohistochemistry

---

1. Introduction

Immunohistochemistry (IHC) enables the visualization of disease-specific protein expression, such as Ki67, PD-L1, and CD3, and plays a vital role in cancer subtyping [1], targeted therapy selection [2, 3, 4], and prognosis assessment [5, 6]. Quantitative interpretation of IHC results is critical for clinical decision-making. For example, the tumor proportion score (TPS) [7], which estimates the fraction of tumor cells expressing a specific biomarker, is widely used to guide treatment for patients above defined thresholds [8]. This requires accurate enumeration of relevant cells to ensure reproducible and objective analysis. Manual counting by pathologists is tedious, time-consuming, and subject to inter-observer variability. The inherent complexity of IHC images, including diverse biomarker staining, high chromogen overlaps, and variations in cellular morphology, density, and spatial expression across tissue regions, further complicates this task. These challenges call for automated cell counting methods that are both accurate and robust under real-world IHC conditions.

Existing cell counting methods can be broadly categorized into detection-based [9, 10, 11, 12, 13, 14, 15] and regression-based methods [16, 17, 18]. Detection-based methods identify individual cell instances to infer counts

but often degrade in densely packed regions and rely on costly instance annotations. In contrast, regression-based methods estimate spatial density maps and integrate them to obtain counts, enabling training with lower-cost point annotations and showing superior performance in overlapping-cell scenario [16]. Despite these strengths, IHC scoring typically requires separate enumeration of biomarker-positive and -negative tumor cells rather than a single aggregated count, while most regression-based models rarely support such end-to-end multi-class counting and the use of foundation models (FMs) within this paradigm remains largely underexplored.

In computational pathology, a growing number of studies [19, 20, 21, 22, 23] have focused on pretraining foundation models to capture the inherent representations of histopathological images. Due to heterogeneity in their architectures, training objectives, and data sources, different FMs often exhibit complementary strengths on specific data [24]. To harness the complementary knowledge of multiple FMs, several works [24, 25, 26, 27] have explored agglomeration strategies that distill the knowledge of multiple FMs into a single student model. Yet, existing approaches typically lack task-awareness or interpretability. They often treat all teachers equally or select them based solely on feature similarity to the student, without explicitly evaluating their task-specific effectiveness. As a result, they fail to provide fine-grained, performance-driven teacher selection during distillation, and this aspect has not yet been investigated in cell counting tasks.

To address the aforementioned limitations, we propose a rank-aware agglomeration framework that distills knowledge from multiple FMs into a compact student, CountIHC, capable of retaining or surpassing the counting performance of the strongest teacher FM, while substantially reducing model complexity. Through agglomeration, CountIHC selectively inherits the complementary and generalizable representations of FMs to handle heterogeneous staining and morphological patterns in IHC. Specifically, a Rank-Aware Teacher Selecting (RATS) strategy is introduced, leveraging an unsupervised global-to-local patch ranking task modeling ordinal relationships in cell quantity. By evaluating teachers on their inherent ability to preserve the correct ranking within each patch group, RATS dynamically selects the optimal teacher. To achieve end-to-end multi-class counting, we design a fine-tuning stage that reformulates counting as vision–language alignment, where discrete semantic anchors derived from structured text prompts encode both category and quantity information. Aligning these anchors with image features regresses class-specific density maps, ensuring accurate esti-

mation even in cell overlapping regions. Our contributions are summarized as follows:

- To the best of our knowledge, it is the first work to apply multiple foundation models agglomeration to IHC cell counting. We propose a rank-aware agglomeration framework that produces a compact model, CountIHC, which retains or surpasses the best-performing teacher FM while substantially reducing model complexity.
- For foundation model agglomeration, we design Rank-Aware Teacher Selecting (RATS), a task-driven teacher selection strategy that employs an unsupervised global-to-local patch ranking task. Aligned with the cell counting objective, it assesses the inherent counting capability of each teacher and enables fine-grained distillation from the optimal one.
- A fine-tuning stage is introduced for multi-class cell counting, where discrete semantic anchors are constructed through structured text prompts to encode both cell category and quantity information, and are aligned with image features to regress class-specific density maps, enhancing inter-class distinction and multi-class counting accuracy.
- Extensive experiments on multi-center IHC datasets with 12 biomarker and 5 tissue types demonstrate the effectiveness of CountIHC, which outperforms state-of-the-art methods and shows strong agreement with pathologists' assessments. The method also exhibits scalability when applied to H&E-stained data.

2. Related work

2.1. Cell counting

Accurate quantification of cells in microscopic images is critically important in both medical and biological domains [28]. In recent years, inspired by advances in counting algorithms from computer vision, deep learning-based cell counting has gained considerable momentum, aiming to alleviate the tedious burden of manual annotation for domain experts. Existing methods can be broadly grouped into two types: detection-based [9, 10, 11, 12, 13, 14, 15] and regression-based methods [16, 17, 18]. Detection-based methods estimate cell counts by identifying individual instances using annotations such as bounding boxes or segmentation masks. Paulauskaite-Taraseviciene et al.

[13], for example, explored the use of Mask R-CNN [14] for overlapping nuclei detection through a two-stage procedure involving potential region proposal and joint mask prediction. Recent methods often incorporate cell-specific proxy maps, such as the vector flow in Cellpose [11] and the distance map in HoVerNet [15]. CellViT [10] further leverages large-scale pretrained models to achieve high-precision instance segmentation of cell nuclei in whole-slide images (WSIs). However, these detection-based methods still face challenges in counting cells that are densely clustered, strongly adhesive, and embedded within complex histopathological structures [16].

Regression-based cell counting methods have attracted increasing attention due to their superior ability to handle overlapping cells and their reliance on low-cost point-level annotations. These methods predict spatial cell density maps and infer the total count by integrating or summing over the entire predicted map. Additionally, the local maxima in the density map can serve as approximations of cell centroids. He et al. [16] extended the FCRN-based density regression framework [17] by incorporating deep supervision to facilitate intermediate feature learning. Zheng et al. [18] proposed a decoupled architecture that separates counting and localization into two dedicated branches and introduces global context modeling, achieving sate-of-the-art performance across multiple cell counting benchmarks. Recently, studies [29, 30, 31] have explored the use of vision-language models such as CLIP [32] for crowd counting, using text prompts to guide density estimation. Although promising in natural scene applications, their applicability to cell or nuclei counting remains underexplored. Moreover, most existing methods do not support simultaneous multi-class cell counting within a single image, limiting their applicability in scenarios where distinct cell categories must be quantified independently.

IHC scoring requires accurate counting of relevant cells or nuclei to assess protein expression, thereby supporting clinical decision-making [9]. Recent efforts have increasingly focused on automating biomarker evaluation using cell-level detection, segmentation, and scoring techniques. For Ki-67, BCData [33] offers a large-scale benchmark dataset that enables proliferation index estimation based on positive and negative tumor cell annotations. DeepLIIF [9] leverages multiplex immunofluorescence as a pixel-level reference to jointly learn cell segmentation and expression prediction. Yan et al. [34] developed a PD-L1 scoring pipeline by integrating established detection and segmentation modules with rule-based metrics to compute clinical indices (e.g., TPS). Brattoli et al. [8] proposed a universal analyzer for HER2,

ER, and PR assessment, employing existing detection models to facilitate cross-marker generalization. While these approaches show promise in specific clinical settings, they are primarily detection-based and thus face challenges in managing densely packed or chromogen-overlapping cells, which are common in IHC images. Moreover, their applicability remains confined to limited biomarker and tissue settings, with insufficient validation on broader and more heterogeneous IHC datasets.

2.2. Multiple foundation models agglomeration

A wide range of foundation models (FMs) have recently emerged in computer vision and related fields. Trained on large-scale datasets, they exhibit strong generalization and broad applicability to diverse downstream tasks. For example, CLIP [32], trained on web-scale image–text pairs, achieves impressive zero-shot performance across visual benchmarks. In computational pathology, self-supervised pretraining methods such as DINO [35] and iBOT [36] have been adopted to build pathology-specific FMs that reach state-of-the-art performance on downstream tasks [21, 22, 23, 37, 38, 39]. One representative example is Virchow2 [21], trained on 3.1 million whole-slide images (WSIs) covering both H&E and IHC stains, which achieves strong results across multiple pathology benchmarks through domain-adapted self-supervised learning.

Although foundation models exhibit broad applicability, training them from scratch remains computationally expensive and time-consuming. Even when adapted to downstream tasks, large-scale variants are often required to sustain performance, creating heavy resource demands that hinder clinical deployment. To mitigate these challenges, knowledge distillation (KD) [40] transfers the representational capacity of a large teacher to a compact student trained on the same domain data. It has been applied to computational pathology foundation models [21, 41], although performance degradation often accompanies size reduction. This gap largely stems from the limited efficacy of conventional feature or label matching strategies [42].

Given the diversity of training data, learning strategies, and downstream performance across foundation models, recent studies [24, 25, 26, 43, 27, 42] have explored agglomerative distillation to unify multiple FMs and harness their complementary strengths for improved generalization. While aggregation, ensemble, and fusion approaches [44, 45, 46, 47, 48] combine intermediate features or decision outputs, agglomeration distills the knowledge of

multiple pretrained teachers into a single model, enabling a unified representation space derived from their learned competencies. AM-RADIO [25] first integrated representations from CLIP [32], DINO [35], and SAM [49], achieving robust performance across heterogeneous vision tasks. RADIOv2.5 [43] further addressed resolution imbalance and token compression to enhance scalability in high-resolution domains. UNIC [26] incorporated ladder projectors and stochastic teacher dropping, resulting in a universal encoder that achieves competitive results across multiple classification benchmarks. Extending this paradigm to computational pathology, GPFM [24] employed unified knowledge distillation from multiple expert encoders, demonstrating improved generalizability across a wide spectrum of clinical applications.

Current agglomeration strategies predominantly fall into two categories: equal learning [24, 25, 43] and teacher dropping regularization (tdrop) [26, 27]. In the former, the contributions of all teacher models are treated equally by averaging their respective distill losses. However, Sarıyıldız et al. [26] observed that teachers do not contribute equally without further intervention. To mitigate this, they introduced a regularization strategy based on feature-level similarity: with the highest loss magnitude is retained during backpropagation, while the others are randomly discarded according to a predefined probability. Although student models derived from these strategies outperform teachers on specific tasks, the absence of explicit teacher capacity evaluation limits their ability to adaptively allocate supervision according to task-specific performance. They operate in a task-agnostic and opaque manner, ignoring model heterogeneity on a per-sample basis.

3. Methods

3.1. Overview

The overview of our proposed method is shown in Fig. 1. It comprises two successive stages: a rank-aware agglomeration stage for task-driven and fine-grained learning from multiple foundation models, and an anchor-guided fine-tuning stage tailored for multi-class cell counting, with the incorporation of the text modality.

During the agglomeration stage, we propose a novel strategy termed Rank-Aware Teacher Selecting (RATS) to identify the optimal teacher foundation model and guide subsequent learning. Specifically, RATS evaluates multiple frozen foundation models on a task-relevant proxy: a global-to-local patch ranking task constructed from spatially ranked cropped regions of each

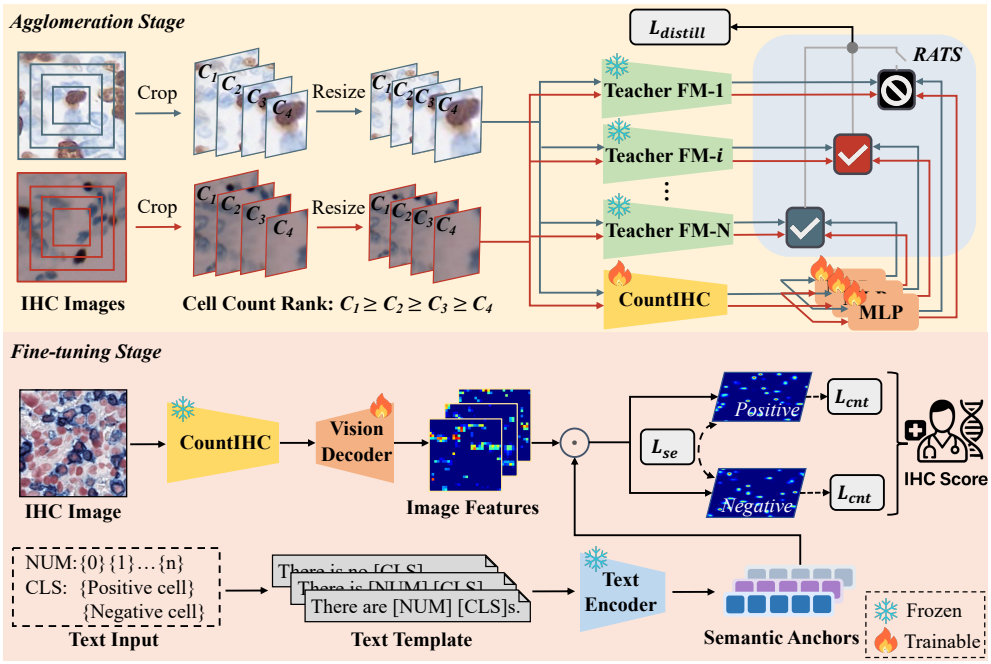

Figure 1: Overview of the proposed rank-aware agglomeration framework and anchor-guided fine-tuning for multi-class cell counting.

input image. Without requiring annotations, this ordinal task provides a meaningful measure of each model's inherent counting ability. RATS selects the optimal teacher sample-wise based on ranking consistency, and routes its knowledge into a compact student model via a distillation loss $L_{distill}$.

In the fine-tuning stage, we introduce discrete semantic anchors that encode both cell quantity ("NUM") and category ("CLS") information through structured text prompts. The alignment between visual features and these anchors enables anchor-guided density map generation across multiple categories. To further enhance the spatial separability of predicted density maps, we propose a Spatial Exclusivity Loss L_{se}, which penalizes inter-class co-activation and strengthens class-wise discriminability. The counts of positive and negative tumor cells derived from the predicted density maps are subsequently utilized for computing IHC scores, such as TPS or Ki-67 labeling index (LI).

3.2. Agglomeration framework: setup and preliminaries

Foundation models trained with different paradigms often exhibit complementary capabilities. For example, CLIP-style vision–language models emphasize semantic abstraction, while DINO-style self-distillation models better preserve spatial and structural fidelity [32, 50]. The goal of the agglomeration framework is to distill knowledge from a pool of N teacher foundation models $\{T_1, T_2, \ldots, T_N\}$ into a single student model S. In this study, we select two representative foundation models from computational pathology, Virchow2 [21] and H-optimus-1 [22], which are among the few trained with extensive IHC-stained WSIs and have demonstrated remarkable performance across various downstream benchmarks. The motivation is to aggregate their complementary representations and inherit their domain-specific capacity to model inherent IHC characteristics. In addition, we incorporate CLIP [32] as a third teacher, inspired by its strong performance in recent crowd counting tasks [29, 30, 31]. Our aim is to extend the paradigm of reformulating counting into a vision-language alignment problem to the IHC multi-class setting, which requires the student model to acquire multimodal alignment capabilities. To balance performance and computational efficiency for practical deployment, we adopt a compact ViT-Base [51] architecture as the student encoder.

To enable feature-level alignment between the student model S and each teacher T_i, we introduce a lightweight teacher-specific projector between S and each T_i. Each projector is implemented as a two-layer multilayer perceptron (MLP) to adapt feature dimensions and facilitate knowledge transfer. Moreover, following findings from prior work [26, 43], we integrate intermediate representations from multiple layers, which have been shown to benefit downstream performance. Specifically, we apply a set of projectors to the outputs of selected intermediate layers and aggregate them to obtain the final output from the student. The alignment of the student and each teacher output, $S(x)$ and $T(x)$, is jointly supervised by cosine similarity and smooth-ℓ_1 loss, as defined in [25], formulated as:

$$\mathcal{L}_{\text{distill}} = \mathcal{L}_{\cos}(S(x), T(x)) + \mathcal{L}_{\text{smooth-}\ell_1}(S(x), T(x)) \tag{1}$$

Regarding the agglomeration strategy, a straightforward way is to treat all teacher models equally by computing $\mathcal{L}_{\text{distill}}$ between the student and each teacher, then summing the losses across all teachers [24, 25, 43]. However, as noted in [26], teachers do not contribute equally without further intervention. To address this, some approaches adopt teacher dropping based on

feature-level similarity [26, 27]: they retain the teacher with the highest distillation loss and randomly discard others, aiming to balance the influence among multiple teachers. Despite this, such selection remains heuristic and lacks interpretability. It offers no principled mechanism for evaluating heterogeneous teachers based on their effectiveness on the target task and data. As a result, it fails to achieve fine-grained and task-specific guidance. This motivates us to design a more informed teacher selection strategy, introduced in the following subsection.

3.3. Rank-Aware Teacher Selecting

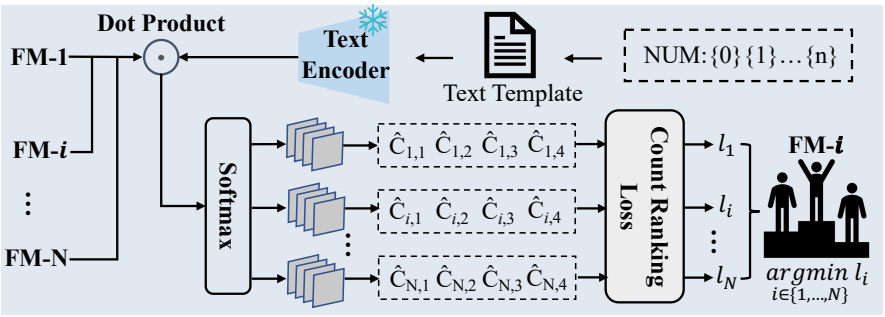

Figure 2: Illustration of the Rank-Aware Teacher Selecting (RATS) strategy. In an unsupervised setting, given a global-to-local cropped patch group, the output of each candidate foundation model (FM-i) is aligned with discrete count candidates to yield a predicted count $\hat{C}_{i,j}$ for patch p_j. A count-ranking loss serves as the rank-aware criterion, and the model that performs best by this criterion is selected as the group-specific teacher.

To provide a unified and label-free measure of each foundation model's counting capacity on the same input, we introduce an unsupervised proxy task: global-to-local patch ranking. This auxiliary task is highly relevant to the underlying goal of cell counting, and enables teacher selection based on each model's inherent ability to preserve ordinal relationships in spatially cropped patch groups.

To construct ordinal supervision without annotation, we leverage the prior that, under center-aligned cropping, larger patches tend to contain equal or greater numbers of cells compared to their nested sub-regions. Based on this observation, we generate ranked patch groups from input IHC images using a global-to-local cropping strategy. Formally, for each image, a sequence of k patches $\{p_1, p_2, \ldots, p_k\}$ is constructed as follows:

1. The largest patch p_k is cropped from raw image using a predefined maximum size M.
2. Centered at the geometric center of p_1, a sequence of progressively smaller patches $\{p_1, \ldots, p_{k-1}\}$ is extracted using a set of scale ratios $\{s_1, \ldots, s_{k-1}\} \subset (0, 1)$, where each patch p_i has size $M \times s_i$. This preserves the aspect ratio and avoids cell deformation or distortion.
3. All cropped patches $\{p_i\}_{i=1}^{k}$ are then resized to a uniform resolution $M \times M$ for model input compatibility.

To estimate the cell quantity in each patch, we sequentially feed the ranked patch group into each of the frozen teacher models. For a given teacher T_i, this produces a sequence of feature embeddings $\{y_{i,1}, y_{i,2}, \ldots, y_{i,k}\}$, where $y_{i,j} = T_i(p_j)$ represents the feature vector extracted from patch p_j. As illustrated in Fig. 2, each embedding is compared against a predefined set of semantic anchors using cosine similarity, and a softmax operation is applied to obtain a probability distribution over discrete count candidates. The predicted count $\hat{C}_{i,j}$ for patch p_j under teacher T_i is then computed as the expectation of this distribution. The construction of semantic anchors and count prediction process will be detailed in subsection 3.4. While the absolute values of predicted counts may be noisy under this unsupervised setting, our focus lies in modeling their ordinal relationship, emphasizing relative ordering rather than exact quantities.

We employ a count ranking loss to evaluate each teacher model's inherent capacity for perceiving ordinal relationships among patches, thereby serving as a unified and interpretable criterion for teacher selection. Given a batch of B patches divided into $G = B/k$ groups, where g-th group consists of a ranked patch list $\{p_1^{(g)}, p_2^{(g)}, \ldots, p_k^{(g)}\}$, the count ranking loss is formulated as:

$$\mathcal{L}_{\text{rank}} = \sum_{g=1}^{G} \sum_{(i,j)\in\mathcal{P}_g} \max\left(0, -(\hat{C}_j^{(g)} - \hat{C}_i^{(g)}) + \varepsilon\right), \tag{2}$$

where $\hat{C}_i^{(g)}$ and $\hat{C}_j^{(g)}$ are the predicted cell counts for patches $p_i^{(g)}$ and $p_j^{(g)}$ within group g, and $\mathcal{P}_g = \{(i, j) \mid 1 \leq i < j \leq k\}$ denotes the set of all ranked patch index pairs within group g. The ranking margin ε is set to zero. This loss penalizes cases where the predicted ranking contradicts the reference order, assuming without loss of generality that $\hat{C}_j \geq \hat{C}_i$. The order follows the predefined global-to-local ranking prior, where patches with larger indices contain equal or larger numbers of cells.

For each teacher T_i, the ranking loss $\mathcal{L}^i_{\text{rank}}$ is computed as above, and the teacher selection follows the rank-aware criterion:

$$i^\star = \underset{i\in\{1,\dots,N\}}{\arg\min}\, \mathcal{L}^i_{\text{rank}}, \qquad T_{\text{selected}} = T_{i^\star}. \tag{3}$$

This selection is performed independently for each training batch, enabling computationally advantageous while allowing fine-grained identification of the foundation model with the optimal inherent counting performance.

3.4. Fine-tuning for multi-class cell counting

To further enhance the accuracy of cell quantification, we fine-tune the distilled student model, CountIHC, which has already acquired strong IHC image representations and an inherent sense of relative cell quantity through rank-aware agglomeration. Considering the clinical requirements of IHC scoring, particularly the need to quantify both positive and negative tumor cells, we design a multi-class cell counting framework driven by semantic anchors, as shown in Fig. 1. It is worth noting that although our experiments primarily focus on these two categories, the framework can be naturally extended to cell counting tasks involving more categories. To reinforce category-level discrimination, we additionally introduce competitive exclusion constraints that penalize spatial overlap between predicted density distributions of different cell categories.

For the IHC image modality, we input the image into the frozen student model CountIHC to extract features. These features are then fed into a trainable vision decoder, implemented as a residual block (ResBlock) [52], which functions as a lightweight adapter for cross-modality feature alignment. The resulting aligned feature map is denoted as $F \in \mathbb{R}^{H'\times W'\times d}$, where $H' = H//p$ and $W' = W//p$, and d is the feature dimension, with p denoting the fixed patch size in the ViT tokenizer [51].

For the text modality, we leverage the discreteness and semantic flexibility of natural language to introduce a set of discrete semantic anchors that encode both cell category and quantity information. This design draws inspiration from blockwise regression strategies in crowd counting [29, 53, 54] and is tailored to the requirements of multi-class cell counting in IHC images.

We begin by defining a category list $\mathcal{Z} = \{z_1, z_2, \dots, z_m\}$, where m denotes the total number of cell categories and z_i represents the i-th category. For IHC images, $\mathcal{Z}$ typically includes two categories: positive tumor cell and negative tumor cell. We then define a discrete count set $\{0, 1, \dots, n\}$, where

n is a truncation threshold introduced to mitigate the influence of annotation noise and imaging artifacts. This set is converted into a collection of count bins $\mathcal{B} = \{\{0\}, \{1\}, \ldots, [n, \infty)\}$. Each category–bin pair $(z_i, b_j) \in \mathcal{Z} \times \mathcal{B}$ is mapped to a natural language prompt using a template function $\Phi(z_i, b_j)$ defined as:

$$\Phi(z_i, b_j) = \begin{cases} \text{“There is no } [z_i]\text{”} & \text{if } b_j = \{0\} \\ \text{“There is } [k]\ [z_i]\text{”} & \text{if } b_j = \{1\} \\ \text{“There are } [k]\ [z_i]\text{s”} & \text{if } b_j = \{k\},\ 2 \leq k < n \\ \text{“There are more than } n\ [z_i]\text{s”} & \text{if } b_j = [n, \infty) \end{cases}$$

All prompts are tokenized using tokenizer of CLIP [32] and encoded by its frozen text encoder, producing the semantic anchor tensor $A \in \mathbb{R}^{m \times (n+1) \times d}$, which incorporates both category and quantity semantics.

Similar to CLIP [32], we compute the cosine similarity between the encoded visual features and each semantic anchor. The resulting similarity scores are normalized via a softmax operation, yielding a probability tensor $P \in \mathbb{R}^{H' \times W' \times m \times (n+1)}$, where each sub-map $P_{i,j}(u, v)$ indicates the likelihood of count bin b_j for category z_i at spatial location (u, v). The final multi-class density maps $D \in \mathbb{R}^{H' \times W' \times m}$ are then computed as the expectation over all count bins:

$$D_i(u, v) = \sum_{j=1}^{n} P_{i,j}(u, v) \cdot b_j \tag{4}$$

where $D_i(u, v)$ denotes the predicted density for category z_i at location (u, v), and b_j is the value of the j-th count bin.

While soft density estimation allows for continuous and overlapping responses, such overlap across semantic channels can be semantically invalid and biologically implausible. To impose inter-class exclusivity and reduce semantic interference, we propose a Spatial Exclusivity Loss (SELoss). SELoss imposes a soft mutual inhibition constraint between class-specific density maps by penalizing their high-confidence co-activation at each pixel. This regularization acts as a differentiable spatial competition mechanism, encouraging the network to resolve ambiguous activations in favor of the most semantically plausible category. Given two predicted density maps D_1, D_2 corresponding to different categories, the SELoss is defined as:

$$\mathcal{L}_{se} = \frac{1}{H'W'} \sum_{u=1}^{H'} \sum_{v=1}^{W'} \sigma_\alpha(D_1(u, v) - \tau) \cdot \sigma_\alpha(D_2(u, v) - \tau)$$

where $\sigma_\alpha(x) = \frac{1}{1+e^{-\alpha x}}$ is a temperature-controlled sigmoid function with steepness parameter α, and τ is a confidence threshold that filters out low-activation noise. The sigmoid-based soft filtering ensures that only confident predictions (above threshold τ) are considered, while α controls the sharpness of exclusivity enforcement.

To further supervise both coarse-grained semantic alignment and fine-grained count accuracy, we adopt the combination of blockwise classification and distribution-matching regression losses as proposed in CLIP-EBC [29], where the regression component is derived from DMCount [55]. Moreover, to address class imbalance, we introduce a category-weighted extension of this objective, resulting in the following category-aware counting loss:

$$\mathcal{L}_{\text{cnt}} = \sum_{i=1}^{m} \lambda_i \left(\mathcal{L}_{\text{ce}}^{(i)} + \mathcal{L}_{\text{dm}}^{(i)} \right) \tag{5}$$

$$\mathcal{L}_{\text{ce}}^{(i)} = -\sum_{u=1}^{H'} \sum_{v=1}^{W'} \sum_{j=1}^{n} \mathbb{1}(P_{i,j}(u,v) = 1) \cdot \log \left(\hat{P}_{i,j}(u,v) \right) \tag{6}$$

$$\mathcal{L}_{\text{dm}}^{(i)} = \left| \|D_i\|_1 - \left\|\hat{D}_i\right\|_1 \right| + \mathcal{W}_2^2 \left(\frac{D_i}{\|D_i\|_1}, \frac{\hat{D}_i}{\left\|\hat{D}_i\right\|_1} \right) + \frac{1}{2} \|D_i\|_1 \left\| \frac{D_i}{\|D_i\|_1} - \frac{\hat{D}_i}{\left\|\hat{D}_i\right\|_1} \right\|_1 \tag{7}$$

where λ_i is a class-specific balancing coefficient for category i. The cross-entropy term $\mathcal{L}_{\text{ce}}^{(i)}$ supervises the predicted probability map $\hat{P}_{i,j}(u,v)$. The indicator function $\mathbb{1}(\cdot)$ is used to focus on the pixels where the ground-truth probability is 1. The distribution matching loss $\mathcal{L}_{\text{dm}}^{(i)}$ consists of three terms: the L1 difference between the total predicted and ground-truth counts, the squared 2-Wasserstein cost $\mathcal{W}_2^2(\cdot,\cdot)$ defined as the optimal transport cost with a quadratic ground metric between the normalized density maps, and an ℓ_1 penalty on their distributional difference. The ground-truth density maps D_i and the predicted counterparts $\hat{D}_i$ are non-negative, and their normalized forms are treated as discrete probability distributions.

Combining the category-aware counting supervision and inter-class spatial exclusivity, the overall training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cnt}} + \gamma \cdot \mathcal{L}_{\text{se}} \tag{8}$$

where γ is a hyperparameter that controls the relative weight of the exclusivity term.

4. Experiment

4.1. Evaluation protocols

4.1.1. Dataset

Table 1: Summary of IHC datasets used in this study. The symbol '#' denotes the number of samples, and 'ATC' refers to annotated tumor cells. The 'PA' denotes whether the dataset is publicly available or not.

Dataset	# Images	# ATC	Resolution	Biomarker(s)	Tissue	PA
Ki67-Camera	1,776	49,124	448×448	Ki-67	Breast	×
Ki67-WSI	11,409	832,493	448×448	Ki-67	Breast	×
IHC-MBM	516	66,196	448×448	CD11b, CD14, CD68, CD163, PD-L1, TIM3, Siglec15, CD73, CD86	Gastric	×
BCData [33]	1,338	181,074	640×640	Ki-67	Breast	✓
DeepLIIF-Data [9]	4,484	98,056	224×224	Ki-67	Lung, Bladder	✓
LyNSeC 1 [56]	1,516	87,316	224×224	CD3, Ki-67, ERG	Lymphoid	✓

As summarized in Table 1, the datasets used in this study span 12 IHC biomarkers and 5 tissue types, including both private collections and publicly available resources. In the agglomeration stage, we utilize two Ki-67–stained private datasets, Ki67-Camera and Ki67-WSI, along with the public LyNSeC 1 [56] dataset for unsupervised training. Subsequently, all datasets are employed to evaluate IHC counting performance, providing a comprehensive basis for assessing the proposed method across varying staining patterns and cell morphologies. Detailed dataset descriptions follow.

IHC-Ki67. We collected two private Ki67-stained datasets with pixel-level annotations, comprising 618,708 negative and 262,909 positive tumor nuclei. These datasets were derived from breast tissue samples of 317 patients across Xiangya Hospital and Quanzhou Hospital. Specifically, **Ki67-Camera** was captured using a 20× microscope; **Ki67-WSI** was acquired as whole-slide images (WSIs) scanned at 20× and 40× magnification. All images were uniformly cropped into non-overlapping patches of size 448×448 pixels.

IHC-MBM. To validate the model's performance across a broader range of biomarker samples, we further collected 20× magnification WSIs from 31 gastric cancer patients at Zhongshan Hospital. The IHC-Multi-BioMarker (IHC-MBM) dataset includes immune markers such as myeloid markers (CD11b, CD14, CD68, CD163), immune checkpoints (PD-L1, TIM3, Siglec15), and co-stimulatory molecules (CD73, CD86). A total of 516 non-overlapping 448×448 patches were cropped from the WSIs.

BCData. The BCData [33] consists of 1,338 Ki-67 IHC image patches (640×640 pixels) extracted from 394 breast cancer WSIs, where each patch was manually annotated by pathologists with the coordinates of all tumor cell centers, which are further labeled as positive or negative based on Ki-67 staining, totally 181,074 point-level annotated tumor cells.

DeepLIIF-Data. We employed the DeepLIIF dataset publicly released by Ghahremani et al. [9], which contains co-registered images of IHC, hematoxylin, and multiplex immunofluorescence channels derived from lung and bladder tissues. For our experiments, we extracted the IHC parts and further cropped it into 4,484 non-overlapping patches of 224×224 pixels.

LyNSeC 1. LyNSeC 1 [56] includes 87,316 labeled cells across three immunomarkers (CD3, Ki-67, and ERG), annotated with nuclear contours and binary cell-level labels indicating marker-positive or marker-negative status. For our experiments, we extracted non-overlapping 224×224 patches from the original images to construct the input set.

4.1.2. Evaluation metrics

To quantitatively evaluate the counting performance for each tumor cell category (positive and negative), we employ the mean absolute error (MAE) and root mean square error (RMSE), defined as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|C_i - \hat{C}_i\right|, \tag{9}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(C_i - \hat{C}_i\right)^2}, \tag{10}$$

where N is the number of images in the test set, C_i is the ground-truth positive/negative tumor cell count value of image i, and $\hat{C}_i$ is the predicted count value.

Furthermore, to evaluate the model's ability to produce clinically meaningful stratification in IHC-based analysis, we adopt the tumor proportion score (TPS), a widely used pathological metric for therapies targeting markers such as PD-L1 and HER2. It is defined as:

$$\mathrm{TPS} = \frac{C_{\mathrm{pos}}}{C_{\mathrm{pos}} + C_{\mathrm{neg}}}, \tag{11}$$

where C_{pos} and C_{neg} denote the number of marker-positive and marker-negative tumor cells, respectively. Based on this definition, we compute the mean absolute error of TPS predictions of each image using the same formulation as previously defined for MAE.

Lastly, we adopt the weighted mean squared error (WMSE) [57] to provide a class-balanced evaluation of model performance. This metric assigns weights to the per-category MSE based on the prevalence of each cell category in the dataset, mitigating the impact of class imbalance. It is defined as:

$$\text{WMSE} = \frac{1}{m} \sum_{i=1}^{m} \omega_i \cdot \text{MSE}_i, \tag{12}$$

$$\omega_i = \frac{\exp(f_i)}{\sum_{j=1}^{m} \exp(f_j)}, \tag{13}$$

$$f_i = \ln\left(\frac{\bar{C}_i}{\text{Median}(\bar{C})}\right), \tag{14}$$

where m is the number of cell categories, and MSE_i denotes the mean squared error for category i. The term ω_i is a softmax-normalized weight assigned to each category, calculated from the log-ratio f_i, where $\bar{C}_i$ represents the total number of ground-truth cells in category i, and $\text{Median}(\bar{C})$ is the median count across all categories.

4.1.3. Implementation details

In the agglomeration stage, a total of $N = 3$ frozen teacher models are employed, specifically Virchow2 [21], H-Optimus-1 [22], and CLIP-L [32]. To align the output representations of each teacher's image encoder with that of the text encoder, we attach an identical ResBlock [52] to each teacher. Each teacher-ResBlock pair is pretrained for 100 epochs under the same setting. This compensates for modality-specific priors in Virchow2 [21] and H-Optimus-1 [22] for IHC, as well as the advantage of CLIP [32] in vision-language alignment.

The student model, CountIHC, adopts the ViT-Base [51] architecture with a patch size of $p = 14$. For each input image, a ranked patch group of size $k = 4$ is generated via center-aligned cropping, with a maximum patch size $M = 224$ and a scale ratio set to $\{\frac{5}{8}, \frac{3}{4}, \frac{7}{8}, 1\}$. The full framework is trained with a batch size of 128 for 200 epochs using the AdamW optimizer. The learning rate is initialized to 1×10^{-3}, decayed to a minimum of 1×10^{-6}

using a cosine annealing schedule, with linear warm-up over the first 10 epochs.

In the fine-tuning stage, the number of count bins was set to $n = 4$. The SELoss was applied with $\alpha = 10$ and $\tau = 0.3$. The category weights for count loss were set to $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$ for positive and negative tumor cells, respectively, and $\gamma = 0.5$ was used to balance the Spatial Exclusivity Loss. We train our models using the Adam optimizer with a batch size of 128 and an initial learning rate of 1×10^{-3}, which is scheduled to decay via cosine annealing.

All experiments were conducted on 4 NVIDIA A800 GPUs. For fair comparison, all compared methods were evaluated under the same dataset protocol and trained with configurations consistent with those reported in their original papers. For detection-based methods, we fine-tuned Cellpose [11] and CellViT [10], while adopting the official checkpoints for DeepLIIF [9] (DeepLIIF_Latest_Model) and μSAM [12] (vit_h_histopathology). As Cellpose and μSAM do not support explicit cell classification, we applied a post-hoc assignment strategy: for each predicted instance, we identified the ground-truth instance with the highest IoU (thresholded at 0.5) across all classes, and assigned the class label corresponding to the best match.

4.2. Comparison on IHC cell counting

We conduct five-fold cross-validation on six datasets spanning 5 tissues and 12 biomarker stainings to compare fine-tuned CountIHC with state-of-the-art regression-based and detection-based counting methods for IHC cell-counting accuracy. Evaluation adopts four primary metrics: NM (MAE of negative tumor cell counts), PM (MAE of positive tumor cell counts), TM (MAE of tumor proportion score, TPS), and WM (WMSE). The averaged results over the five splits are summarized in Table 2 and Table 3.

Across the two IHC-Ki67 datasets we collected, CountIHC consistently demonstrates advantages under practical IHC conditions spanning different acquisition modalities. As reported in Table 2, for the class-balanced objective WM, CountIHC reduces error relative to the strongest regression baseline (CLIP-EBC [29]) by about 4.9% on Ki67-Camera (6.1 to 5.8) and about 7.8% on Ki67-WSI (23.1 to 21.3). For the tumor proportion score, which is clinically important because it quantifies tumor-cell expression and guides subsequent treatment decisions, countIHC achieves lower TM than CLIP-EBC [29] in Ki67 settings. Unlike CLIP-EBC [29], which relies on visual-prompt tuning of CLIP [32] for downstream adaptation, our approach adds

Table 2: Quantitative results of IHC cell counting on three private datasets using five-fold cross-validation. Reg-based and Det-based denote regression-based and detection-based methods, respectively. Evaluation metrics include NM (MAE of negative tumor cell counts), PM (MAE of positive tumor cell counts), TM (MAE of tumor proportion score, TPS, %), and WM (WMSE). The downward arrow (↓) indicates that lower values correspond to better performance. All values are reported as the mean across five folds. **Bold** numbers denote the best results, and underlined numbers indicate the second-best results. Best/second-best results are determined before rounding; although some rounded values appear equal, comparisons are based on the exact numbers.

Model	Type	Ki67-Camera				Ki67-WSI				IHC-MBM			
		NM ↓	PM ↓	TM ↓	WM ↓	NM ↓	PM ↓	TM ↓	WM ↓	NM ↓	PM ↓	TM ↓	WM ↓
CSRNet [58]	Reg-based	3.2	**1.0**	8.5	7.3	9.9	2.7	3.8	52.2	12.6	6.0	5.0	85.0
C-FCRN+Aux [16]		5.1	1.9	11.9	15.5	12.0	3.4	4.3	72.1	21.3	6.6	6.4	126.4
DCL [18]		3.6	1.4	12.7	8.6	7.6	2.3	3.5	28.0	12.1	5.3	**4.1**	74.0
CLIP-EBC [29]		3.1	1.2	7.1	6.1	7.1	2.1	3.3	24.0	12.3	6.5	5.3	89.2
DeepLIIF [9]	Det-based	28.1	6.4	14.9	392.1	42.6	6.2	6.0	431.3	38.0	16.3	9.2	1067.8
CellViT [10]		7.0	1.9	9.0	30.7	17.4	3.7	4.6	134.1	15.5	10.4	6.0	208.5
Cellpose [11]		5.4	2.3	7.2	21.9	18.2	4.3	4.5	143.4	22.4	8.7	5.7	161.6
μSAM [12]		5.1	5.9	22.3	45.8	21.0	14.6	18.7	376.6	16.9	11.0	9.1	184.1
CountIHC (Ours)	Reg-based	**3.1**	1.2	**6.9**	**5.8**	**6.6**	**2.1**	**3.2**	**21.3**	**10.4**	**5.1**	4.4	**58.4**

no extra computation overhead and performs end-to-end multi-class counting with smaller errors, indicating the effectiveness of the proposed rank-aware agglomeration strategy (RATS) together with the carefully designed multi-class fine-tuning. This advantage in balanced counting remains robust on the broader biomarker dataset (IHC-MBM) in Table 2: CountIHC lowers WM from 74.0 to 58.4 (about 21.1%) while attaining the second-best TM of 4.4. Notably, on IHC-MBM our method achieves the lowest NM (10.4) and PM (5.1) among all compared methods. On the three public IHC datasets, as reported in Table 3, CountIHC also delivers consistent gains in WM, with reductions of approximately 3.6% on DeepLIIF-Data (5.5 to 5.3), 2.7% on LyNSeC-1 (11.0 to 10.7), and 10.2% on BCData (197.7 to 177.6) over the best-performing regression baselines, while keeping TM competitive and generally lowering positive- and negative-cell MAE. In contrast, several baselines show asymmetric behavior between negative and positive tumor cells, improving one at the expense of the other, which leads to larger WM. By improving the aggregate error without favoring either class, CountIHC shows reliable performance across diverse biomarker stainings.

Table 3: Quantitative results of IHC cell counting on three public datasets using five-fold cross-validation. Evaluation metrics include NM (MAE of negative tumor cell counts), PM (MAE of positive tumor cell counts), TM (MAE of tumor proportion score, TPS, %), and WM (WMSE). The downward arrow (↓) indicates that lower values correspond to better performance. All values are reported as the mean across five folds. **Bold** numbers denote the best results, and underlined numbers indicate the second-best results. Best/second-best results are determined before rounding; although some rounded values appear equal, comparisons are based on the exact numbers.

Model	DeepLIIF-Data				LyNSeC 1				BCData			
	NM ↓	PM ↓	TM ↓	WM ↓	NM ↓	PM ↓	TM ↓	WM ↓	NM ↓	PM ↓	TM ↓	WM ↓
CSRNet [58]	**2.2**	1.7	8.0	5.5	3.2	2.8	4.8	14.3	17.5	8.2	**3.9**	197.7
C-FCRN+Aux [16]	4.7	3.5	11.9	15.3	5.2	4.1	4.4	26.2	20.7	9.6	6.8	308.3
DCL [18]	2.4	1.7	**7.1**	5.5	3.5	3.2	5.2	17.6	20.2	8.7	5.1	277.8
CLIP-EBC [29]	2.5	2.0	7.4	6.8	2.9	2.4	3.6	11.0	16.4	8.8	5.3	204.0
DeepLIIF [9]	7.8	3.1	11.5	52.3	21.3	34.0	20.2	1206.1	–	–	–	–
CellViT [10]	4.6	4.1	12.4	30.4	3.2	3.8	5.3	20.7	–	–	–	–
Cellpose [11]	8.2	3.8	9.1	20.5	4.7	3.9	**2.7**	20.8	–	–	–	–
μSAM [12]	9.3	6.5	20.9	37.2	27.8	8.6	53.3	369.6	–	–	–	–
CountIHC (Ours)	2.7	**1.7**	7.2	**5.3**	**2.7**	**2.3**	3.6	**10.7**	**15.5**	**7.4**	4.3	**177.6**

Fig. 3 demonstrates qualitative comparisons for the top-performing methods. For each patch, we show density maps from regression-based methods and instance contours from detection-based methods, together with the predicted negative and positive counts. CountIHC produces the closest results to the ground truth, for example, 89.9/5.6 vs. 90/6 and 88.6/19.4 vs. 91/19 (negative / positive) whereas CLIP-EBC [29] underestimates negatives and DCL exhibits noisy responses that bias the negative-positive ratio. Detection-based approaches struggle with overlapped, small, and weakly stained cells, leading to missed counts or spurious regions. The density maps from CountIHC display well-localized peaks with clear separation between the positive and negative channels, consistent with the semantic-anchor design and the Spatial Exclusivity Loss.

4.3. Human-machine agreement comparison

To validate the agreement between the diagnostic outcomes of CountIHC and the clinical assessments made by pathologists, we further evaluated on two independent datasets, Ki67-Camera and Ki67-WSI, following the same source and naming convention as those used in the main experiments. Im-

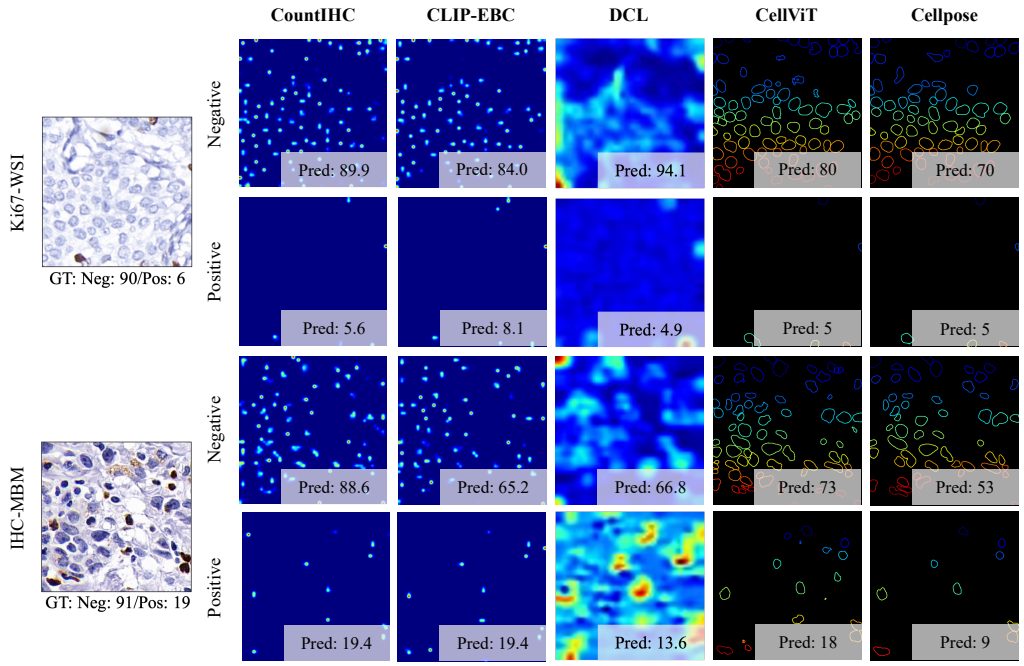

Figure 3: Qualitative comparisons of negative/positive tumor cell counting predictions on IHC images. Ground-truth counts (GT) are shown alongside the predicted counts (Pred) from different methods.

portantly, these datasets are completely separate from those used in previous experiments, thus ensuring an objective and fair comparative evaluation. Each dataset comprises cellular images from 100 patients. For diagnostic reliability, cellular patch images corresponding to each patient were organized into individual folders, with Ki67 assessments provided on the basis of 1–10 representative images per folder.

During the diagnostic process, five experienced pathologists from different hospitals participated in this study, including the First Affiliated Hospital of Xi'an Jiaotong University, Peking University People's Hospital, the Second Affiliated Hospital of Xi'an Jiaotong University, and Quanzhou Hospital. Each pathologist independently reviewed both datasets and provided diagnostic results categorized as Ki67-, Ki67+, Ki67++, and Ki67+++. These categories correspond to increasing levels of the Ki67 labeling index (LI), which is the standard clinical measure for Ki67 expression. The LI is calculated in a manner analogous to the tumor proportion score (TPS), representing the percentage of Ki67-positive tumor cells relative to the total tumor

cell population.

For the machine-based evaluation, we applied the models previously fine-tuned on Ki67-Camera and Ki67-WSI datasets to the corresponding independent test sets, Ki67-Camera and Ki67-WSI, respectively. The inference results yielded the numbers of positive and negative tumor cells, which were subsequently aggregated and used as the machine's reference diagnostic outcomes.

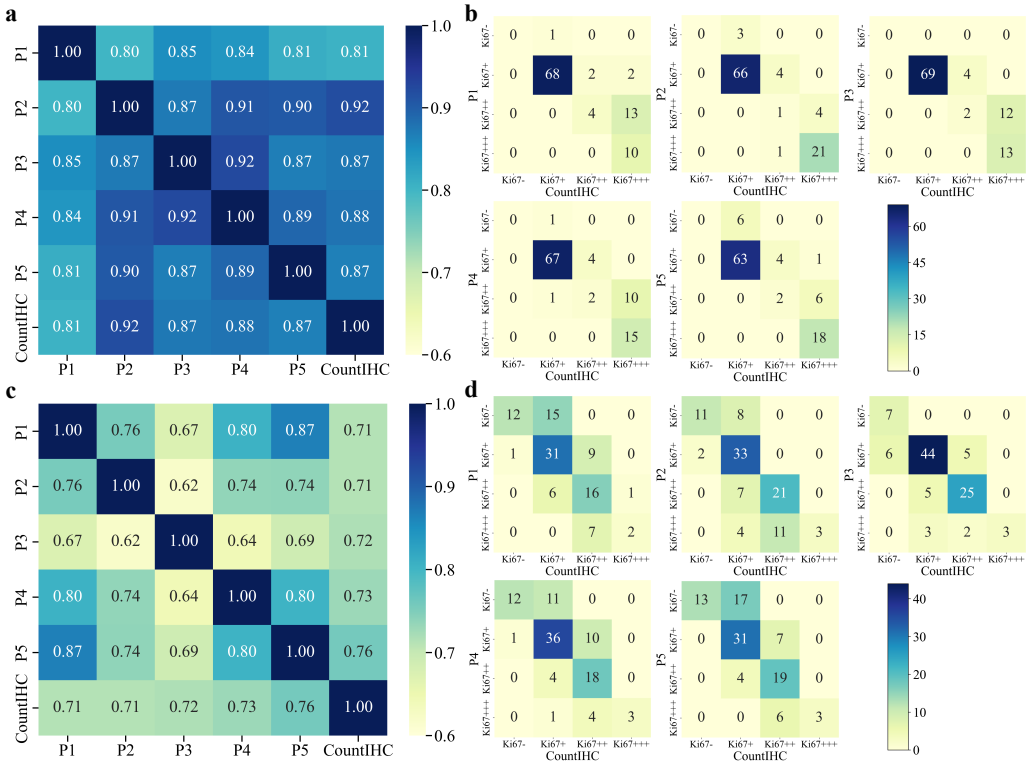

Figure 4: Agreement between the CountIHC and five experienced pathologists (P1–P5) on the Ki67-Camera and Ki67-WSI datasets. Panels **a** and **b** show the results on the Ki67-Camera dataset, with **a** presenting pairwise quadratic weighted kappa (QWK) values between the machine and pathologists and among pathologists, and **b** showing confusion matrices for grade distributions between the machine and each pathologist. Panels **c** and **d** show the corresponding analyses on the Ki67-WSI dataset.

The agreement between machine and pathologists was first assessed using quadratic weighted kappa (QWK), a widely adopted metric for evaluating ordinal-scale agreement. As illustrated in Fig. 4, the QWK heatmaps

demonstrate that on Ki67-Camera, the agreement between the machine and pathologists ranged from 0.81 to 0.92, a level closely comparable to the inter-pathologist agreement (0.80–0.92). This indicates that the model achieved human-level diagnostic reliability on camera-acquired images. On Ki67-WSI, the QWK values between the machine and pathologists were moderately lower, ranging from 0.71 to 0.76, but still within the broader range of inter-pathologist variability (0.62–0.87). These findings suggest that although the diagnostic task becomes more challenging at various magnifications, the machine remains comparable to human experts in terms of diagnostic agreement.

As a further assessment of diagnostic performance, confusion matrices between the machine and each pathologist are shown in Fig. 4. On Ki67-Camera, most cases clustered along the diagonal, indicating strong agreement across all four Ki67 expression categories. Misclassifications, when present, primarily occurred between adjacent categories (e.g., Ki67+ vs. Ki67++), which correspond to subjective boundary regions in clinical practice. On Ki67-WSI, diagonal dominance persisted but off-diagonal entries were more frequent compared with the camera dataset, particularly in intermediate categories. Notably, these discrepancies coincided with regions of reduced inter-pathologist agreement, suggesting that the machine does not introduce systematic biases but rather reflects the inherent variability among human experts. As shown in Fig. 5, we further visualize local maxima of the predicted negative/positive density maps as point markers approximating cell centroids on the tissue image. This qualitative evidence substantiates the reliability of the inferred positive/negative tumor cell distributions.

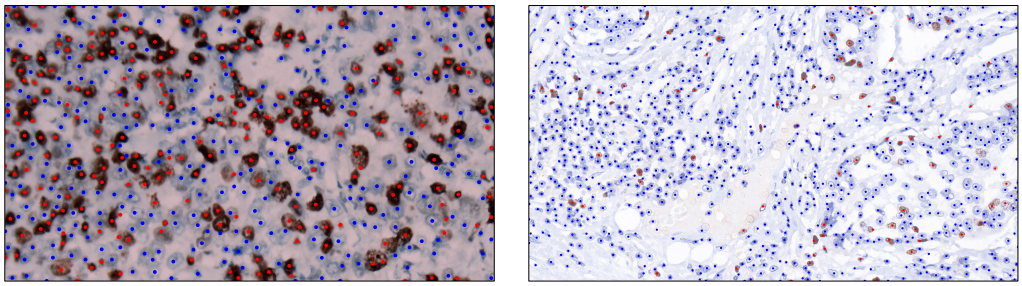

Figure 5: Visualization of approximate cell centroids derived from local maxima of the predicted class-specific density maps on Ki67-Camera and Ki67-WSI. The derived points are overlaid as markers on the tissue images, with **blue** indicating negative tumor cells and **red** indicating positive tumor cells.

These findings highlight that our proposed CountIHC achieves diagnostic performance highly consistent with that of experienced pathologists across different imaging modalities. This evidence supports the system's potential utility as a reliable auxiliary tool in diagnostic pathology, capable of generalizing across acquisition conditions and reducing inter-observer variability in IHC scoring.

4.4. Ablation study

To assess whether CountIHC can inherit or even improve upon the counting ability of teacher foundation models, we use each teacher as the image encoder within our multi-class fine-tuning framework and keep the training and evaluation protocol identical to the main experiments. Without loss of generality, we evaluate on the first fold of four datasets, and we report two primary metrics that capture clinical and class-balanced accuracy: TM (MAE of the tumor proportion score) and WM (weighted mean squared error). The comparisons between the agglomerated student and Virchow2 [21], H-optimus-1 [22], and CLIP-L [32] are summarized in Table 4.

With parameters far fewer than any teacher, CountIHC retain or even surpass the counting performance of the best-performing teachers. On Ki67-Camera and Ki67-WSI it attains the lowest TM and WM, with WM reduced by about 9.2% and 26.4% relative to the strongest teacher. On the multi-biomarker IHC-MBM dataset it achieves the lowest WM, 74.2 vs. 77.6, while keeping TM competitive at 4.1, indicating better balance between negative and positive tumor cell counts under heterogeneous stainings. On BCData the student reaches the best TM of 2.8 with WM closest to the best teacher. Notably, Table 4 also shows that the three teachers exhibit complementary strengths across the four diverse datasets. The proposed Rank-Aware Teacher Selecting (RATS) leverages this complementarity to guide agglomeration at fine granularity, producing a compact student that retains complementary teacher strengths while markedly reducing computational cost, which is advantageous for clinical deployment.

On top of CountIHC, we ablate the effect of the proposed agglomeration framework and the designed fine-tuning for multi-class cell counting on the first fold of the Ki67-WSI dataset, which involve the agglomeration strategy, i.e., Rank-Aware Teacher Selecting (RATS), text-prompt semantic anchors, and the Spatial Exclusivity Loss (SELoss). The setting follows the main training and evaluation protocol. We report the previously defined metrics

Table 4: Quantitative results of the agglomerated student model (CountIHC) and three teacher FMs on the first fold of each dataset. Evaluation metrics include TM (MAE of tumor proportion score, TPS, %), and WM (WMSE). The downward arrow (↓) indicates that lower values correspond to better performance. **Bold** numbers denote the best results, and underlined numbers indicate the second-best results. Best/second-best results are determined before rounding; although some rounded values appear equal, comparisons are based on the exact numbers.

Model	Params ↓ (M)	Ki67-Camera		Ki67-WSI		IHC-MBM		BCData	
		TM ↓	WM ↓	TM ↓	WM ↓	TM ↓	WM ↓	TM ↓	WM ↓
Virchow2 [21]	631.2	6.8	8.7	4.6	62.7	4.4	98.6	3.3	150.0
H-optimus-1 [22]	1134.8	6.6	9.6	3.6	49.6	**3.2**	93.7	2.9	**116.6**
CLIP-L [32]	304.0	9.4	23.2	3.0	42.8	3.73	77.6	4.4	207.2
CountIHC (Ours)	**85.7**	**6.1**	**7.9**	**2.5**	**31.5**	4.1	**74.2**	**2.8**	123.7

and additionally include NR and PR, the root-mean-square errors of negative and positive tumor cell counts. Results are summarized in Table 5.

For the agglomeration stage, we compare RATS with two widely known agglomeration strategies, equal learning [24, 25, 43] and teacher dropping regularization (tdrop) [26, 27], using the same set of teachers and fine-tuning settings. Compared with these strategies, RATS reduces the WM from 48.1 and 36.3 to 31.5 and lowers the TM from 4.3 and 3.6 to 2.5, accompanied by consistent reductions in NM, PM, NR, and PR. We attribute this to a key limitation of these strategies: the absence of an explicit assessment of teacher capacity prevents them from adaptively allocating supervision according to task-specific performance. In contrast, RATS performs task-driven agglomeration by dynamically identifying the optimal teacher based on its inherent counting ability, thereby assigning more appropriate supervision and achieving improved performance.

For semantic anchors, we compare text prompts and rank prompts at both the agglomeration (AG) and fine-tuning (FT) stages. Building on [59], which emphasizes a rank-prompt formulation, we adapt the method by introducing two class-specific rank prompters (one for negative and one for positive tumor cells), while keeping all other settings identical to the optimal configuration reported therein. As shown in Table 5, replacing text prompts with rank prompts at either AG or FT yields inferior results. Using text prompts at both stages provides a clear, class-aware alignment signal without introducing additional modules and achieves the strongest performance, lowering the WM

Table 5: Ablation study on the first fold of Ki67-WSI. AG denotes the agglomeration stage and FT denotes the fine-tuning stage; TP and RP refer to the use of text prompt and rank prompt, respectively. Evaluation metrics include NM (MAE of negative tumor cell counts), NR (RMSE of negative tumor cell counts), PM (MAE of positive tumor cell counts), PR (RMSE of positive tumor cell counts), TM (MAE of tumor proportion score, TPS, %), and WM (weighted mean squared error, WMSE). The downward arrow (↓) indicates that lower values correspond to better performance. **Bold** numbers denote the best results, and underlined numbers indicate the second-best results. w/ and w/o denote "with" and "without," respectively. Best/second-best results are determined before rounding; although some rounded values appear equal, comparisons are based on the exact numbers.

Setting		Ki67-WSI					
		NM ↓	NR ↓	PM ↓	PR ↓	TM ↓	WM ↓
Agglomeration strategy	Equal learning [24, 25, 43]	8.0	12.0	4.3	5.7	4.3	48.1
	tdrop [26, 27]	7.9	12.3	3.4	5.1	3.6	36.3
Prompt	AG-TP&FT-RP	7.5	11.8	2.9	4.6	2.7	34.1
	AG-RP&FT-RP	7.8	11.9	2.9	4.4	2.7	34.1
Loss	w/o SELoss	7.4	11.9	2.7	4.3	2.5	35.4
Ours	RATS AG-TP+FT-TP w/ SELoss	**7.4**	**11.7**	**2.6**	**4.1**	**2.5**	**31.5**

from 34.1 to 31.5 and improving the TM from 2.7 to 2.5. The introduction of SELoss further brings a consistent improvement, reducing the WM from 35.4 to 31.5, with concurrent decreases in NR and PR. By penalizing high-confidence co-activation between class-specific density maps at each pixel, SELoss mitigates cross-class confusion and improves both TPS accuracy and the overall balanced error. As shown in Fig. 6, training without SELoss leads to erroneous co-activation of the negative and positive channels at the same locations, whereas SELoss suppresses this cross-class co-activation and yields sharper, better-separated positive responses at the same locations highlighted with arrows, confirming the intended effect of the loss.

4.5. Applicability to H&E-stained images

To further validate the scalability of our proposed agglomeration framework and fine-tuning strategy, we extended the experiments to H&E-stained images, another widely used modality in clinical pathology. During the agglomeration stage, we adopted seven publicly available datasets (CoNSeP [15], CPM17 [60], LyNSec 2&3 [56], MoNuSeg [61], PanNuke [62], NuInsSeg

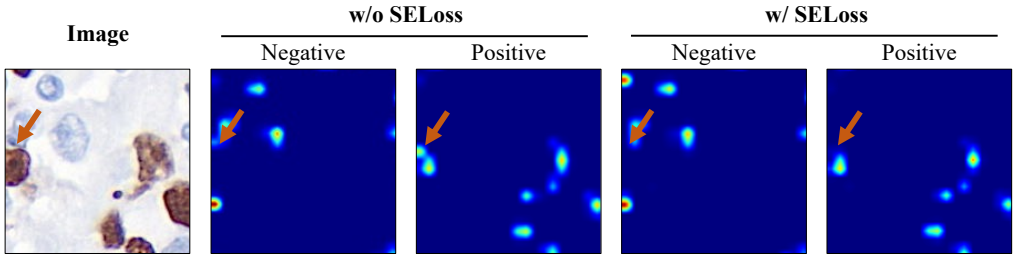

Figure 6: Visualization of SELoss effects on class-specific density maps. Without SELoss, the negative and positive channels spuriously co-activate at the same locations. With SELoss, this cross-class co-activation is suppressed and the positive responses at the arrowed locations are sharper and better separated.

[63], and PUMA [64]) to conduct rank-aware agglomeration of foundation models, in which the same three foundation models (Virchow2 [21], H-optimus-1 [22] and CLIP-L [32]) were distilled into a single student model without relying on manual annotations. The data preprocessing pipeline, methodological settings, and training procedures were kept consistent with those described for IHC-stained images. At the fine-tuning stage, we employed MoNuSeg [61] (cropped into non-overlapping 224×224 patches) and TNBC [65] (cropped into non-overlapping 448×448 patches) to adapt the student model, and subsequently evaluated its cell counting performance on H&E-stained images.

Quantitative results are presented in Table 6. On MoNuSeg, CountIHC achieved the lowest MAE of 2.7 and RMSE of 3.4, improving over the compared method Cellpose (MAE 3.0 / RMSE 3.8) by 10.0% and 10.5%, respectively. On TNBC, CountIHC reached an MAE of 5.0 and RMSE of 7.8, which are competitive with H-optimus-1 [22] (MAE 5.5 / RMSE 6.7), the best-performing teacher FM, while requiring significantly less computation overhead. These results echo the findings on IHC-stained images, demonstrating that our approach consistently retains or even surpasses the counting performance of the best-performing teachers across diverse histopathological modalities.

The qualitative comparisons in Fig. 7 further support these findings. CountIHC yields estimates that remain closely aligned with the ground-truth values (e.g., Pred: 46.1 vs. GT: 50 on MoNuSeg, Pred: 81.5 vs. GT: 81 on TNBC). In contrast, detection-based methods (e.g., CellViT, Cellpose) tend to under-segment or miss densely packed nuclei, while regression-based mod-

Table 6: Quantitative results of cell counting on H&E-stained datasets (MoNuSeg and TNBC). Evaluation metrics include mean absolute error (MAE) and root mean squared error (RMSE). The downward arrow (↓) indicates that lower values correspond to better performance. **Bold** numbers denote the best results, and underlined numbers indicate the second-best results.

Model	MoNuSeg		TNBC	
	MAE ↓	RMSE ↓	MAE ↓	RMSE ↓
Virchow2 [21]	3.1	4.0	7.8	8.9
H-optimus-1 [22]	3.1	3.9	5.5	**6.7**
CLIP-L [32]	3.5	4.7	6.6	8.2
CSRNet [58]	3.2	4.4	7.0	8.1
C-FCRN+Aux [16]	5.1	6.5	12.1	14.2
DCL [18]	4.0	5.2	11.5	12.3
CLIP-EBC [29]	3.3	4.4	15.2	17.1
CellViT [10]	3.8	5.1	16.1	19.6
Cellpose [11]	3.0	3.8	11.7	13.4
CountIHC (Ours)	**2.7**	**3.4**	**5.0**	7.8

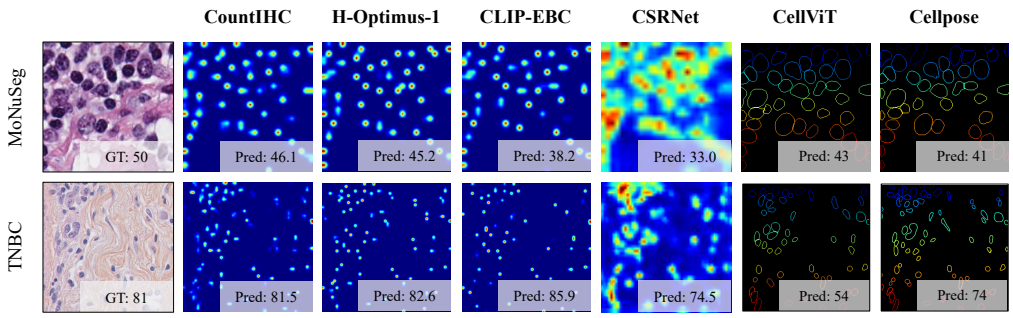

Figure 7: Qualitative comparisons of cell counting predictions on H&E-stained images from MoNuSeg (top) and TNBC (bottom). Ground-truth counts (GT) are shown alongside the predicted counts (Pred) from different methods.

els (e.g., DCL, CSRNet) often generate overly smoothed or noisy predictions, leading to significant under- or over-estimation.

Both quantitative and qualitative analyses verify that the proposed agglomeration framework and fine-tuning strategy extend robustly from IHC- to H&E-stained images, underscoring their scalability for reliable cell counting across modalities.

5. Discussion

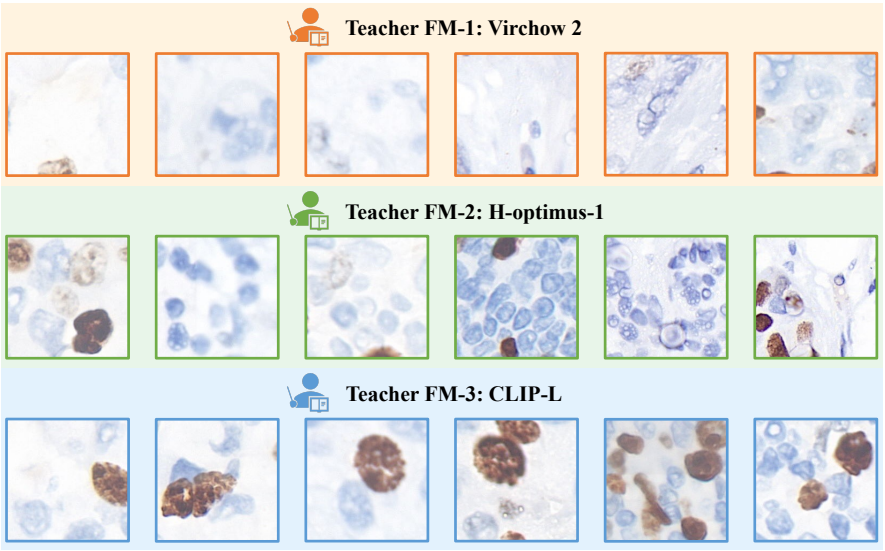

Figure 8: Visualization of sample-level teacher FM assignments by RATS.

Our work presents a rank-aware agglomeration framework that distills complementary strengths from multiple foundation models into a compact student, CountIHC. To further illustrate the complementary counting capacities of different teachers in IHC, Fig. 8 visualizes sample-level teacher assignments produced by the proposed Rank-Aware Teacher Selecting (RATS) strategy. In practice, we observe that H-Optimus-1 [22] is the most frequently selected teacher, particularly for samples with crowded nuclei, pronounced morphological heterogeneity, or weak and uneven staining. Its larger model size and broader pretraining data scale provide more robust supervision, consistent with its role as the overall best-performing teacher reported in Table 5. Virchow2 [21] is more often selected in low-contrast or sparsely cellular regions, where conservative responses help suppress spurious positives and stabilize the negative-cell channel. CLIP-L [32] is preferentially assigned to patches containing prototypical, sharply stained positive tumor nuclei with clear boundaries, where precise localization is critical. The observed teacher-specific assignments arise from the RATS strategy, which employs an unsupervised global-to-local patch ranking task aligned with cell counting. It distills knowledge from the teacher exhibiting higher rank consistency within each patch group, as this indicates stronger inherent counting capability.

Collectively, the results in Fig. 8 confirm that RATS adaptively leverages the complementary strengths of different teachers, allowing CountIHC to remain compact while retaining or even surpassing the performance of the best single teacher.

Combined with semantic anchors that encode both cell quantity and category information through structured text prompts, and a Spatial Exclusivity Loss (SELoss) that enforces inter-class spatial separability, the framework aligns visual features with discrete semantic representations, enabling accurate and category-aware density estimation. Across 6 multi-center IHC datasets covering 12 biomarker stainings and 5 tissue types, five-fold cross-validation against state-of-the-art detection- and regression-based methods yields consistently superior performance on most metrics. A human–machine agreement study further shows a high degree of agreement with pathologists' assessments, underscoring reliability and potential clinical value. Ablations verify that both the proposed agglomeration and the fine-tuning strategy contribute materially to the gains. Finally, transferring the method to H&E-stained data indicates encouraging modality robustness and scalability.

Looking ahead, we plan to scale the data used during agglomeration and extend evaluation beyond patch-level counting to slide- and region-level tasks that are common in IHC practice and benchmarks. These include biomarker scoring such as H-score estimation or PD-L1 CPS, immune contexture analysis such as compartment-specific TIL density, cell phenotyping or subtype classification, and fine-grained instance-level cell segmentation. These studies will further characterize the breadth of capabilities inherited through rank-aware multi-teacher distillation and assess downstream clinical utility.

6. Conclusion

In this work, we present CountIHC, a model derived from rank-aware agglomeration of multiple foundation models, achieving comparable or superior performance to the best teacher FM in IHC cell counting, with substantially reduced computation overhead. Building on the complementary representations of foundation models, CountIHC effectively handles heterogeneous staining and morphological patterns in IHC. Specifically, through the Rank-Aware Teacher Selecting (RATS) strategy, CountIHC dynamically selects the optimal teacher sample-wise via an unsupervised patch ranking task, facilitating adaptive and task-driven knowledge learning. Additionally, CountIHC undergoes a fine-tuning stage that reformulates counting as vision–language

alignment. In this stage, discrete semantic anchors, constructed through structured text prompts, encode both cell category and quantity information. Their alignment with image features regresses class-specific density maps and improves counting accuracy in cell overlapping regions. Extensive experiments across diverse IHC datasets demonstrate that CountIHC outperforms state-of-the-art methods and achieves high agreement with pathologists' assessments, highlighting its potential clinical value. Our method also demonstrates its scalability when applied to H&E-stained data.

Declaration of Generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the authors used ChatGPT for language polishing and grammar refinement only. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] K. Inamura, Update on immunohistochemistry for the diagnosis of lung cancer, Cancers 10 (3) (2018) 72.

[2] M.-K. Song, B.-B. Park, J. Uhm, Understanding immune evasion and therapeutic targeting associated with pd-1/pd-l1 pathway in diffuse large b-cell lymphoma, International journal of molecular sciences 20 (6) (2019) 1326.

[3] T. Zhang, H. Liu, L. Jiao, Z. Zhang, J. He, L. Li, L. Qiu, Z. Qian, S. Zhou, W. Gong, et al., Genetic characteristics involving the pd-1/pd-l1/l2 and cd73/a2ar axes and the immunosuppressive microenvironment in dlbcl, Journal for Immunotherapy of Cancer 10 (4) (2022) e004114.

[4] C. Casulo, A. Santoro, K. Ando, S. Le Gouill, J. Ruan, J. Radford, L. Arcaini, A. Pinto, R. Bouabdallah, K. Izutsu, et al., Durvalumab (anti pd-l1) as monotherapy or in combination therapy for relapsed/refractory (r/r) diffuse large b-cell lymphoma (dlbcl) and follicular lymphoma (fl): a subgroup analysis from the phase 1/2 fusion nhl-001 global multicenter trial, Blood 134 (2019) 5320.

[5] P.-p. Xu, C. Sun, X. Cao, X. Zhao, H.-j. Dai, S. Lu, J.-j. Guo, S.-j. Fu, Y.-x. Liu, S.-c. Li, et al., Immune characteristics of chinese diffuse large b-cell lymphoma patients: implications for cancer immunotherapies, EBioMedicine 33 (2018) 94–104.

[6] L. Qiu, H. Zheng, X. Zhao, The prognostic and clinicopathological significance of pd-l1 expression in patients with diffuse large b-cell lymphoma: a meta-analysis, Bmc Cancer 19 (1) (2019) 273.

[7] M. Boyer, M. A. Şendur, D. Rodríguez-Abreu, K. Park, D. H. Lee, I. Çiçin, P. F. Yumuk, F. J. Orlandi, T. A. Leal, O. Molinier, et al., Pembrolizumab plus ipilimumab or placebo for metastatic non–small-cell lung cancer with pd-l1 tumor proportion score$\geq$ 50%: randomized, double-blind phase iii keynote-598 study, Journal of clinical oncology 39 (21) (2021) 2327–2338.

[8] B. Brattoli, M. Mostafavi, T. Lee, W. Jung, J. Ryu, S. Park, J. Park, S. Pereira, S. Shin, S. Choi, et al., A universal immunohistochemistry analyzer for generalizing ai-driven assessment of immunohistochemistry across immunostains and cancer types, npj Precision Oncology 8 (1) (2024) 277.

[9] P. Ghahremani, Y. Li, A. Kaufman, R. Vanguri, N. Greenwald, M. Angelo, T. J. Hollmann, S. Nadeem, Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification, Nature machine intelligence 4 (4) (2022) 401–412.

[10] F. Hörst, M. Rempe, L. Heine, C. Seibold, J. Keyl, G. Baldini, S. Ugurel, J. Siveke, B. Grünwald, J. Egger, et al., Cellvit: Vision transformers for precise cell segmentation and classification, Medical Image Analysis 94 (2024) 103143.

[11] C. Stringer, M. Pachitariu, Cellpose3: one-click image restoration for improved cellular segmentation, Nature methods 22 (3) (2025) 592–599.

[12] A. Archit, L. Freckmann, S. Nair, N. Khalid, P. Hilt, V. Rajashekar, M. Freitag, C. Teuber, M. Spitzner, C. Tapia Contreras, et al., Segment anything for microscopy, Nature Methods 22 (3) (2025) 579–591.

[13] A. Paulauskaite-Taraseviciene, K. Sutiene, J. Valotka, V. Raudonis, T. Iesmantas, Deep learning-based detection of overlapping cells, in:

Proceedings of the 3rd International Conference on Advances in Artificial Intelligence, 2019, pp. 217–220.

[14] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[15] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, N. Rajpoot, Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images, Medical image analysis 58 (2019) 101563.

[16] S. He, K. T. Minn, L. Solnica-Krezel, M. A. Anastasio, H. Li, Deeply-supervised density regression for automatic cell counting in microscopy images, Medical Image Analysis 68 (2021) 101892.

[17] W. Xie, J. A. Noble, A. Zisserman, Microscopy cell counting and detection with fully convolutional regression networks, Computer methods in biomechanics and biomedical engineering: Imaging & Visualization 6 (3) (2018) 283–292.

[18] Z. Zheng, Y. Shi, C. Li, J. Hu, X. X. Zhu, L. Mou, Rethinking cell counting methods: Decoupling counting and localization, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 418–426.

[19] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, X. Han, Transformer-based unsupervised contrastive learning for histopathological image classification, Medical image analysis 81 (2022) 102559.

[20] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, J. Zou, A visual–language foundation model for pathology image analysis using medical twitter, Nature medicine 29 (9) (2023) 2307–2316.

[21] E. Zimmermann, E. Vorontsov, J. Viret, A. Casson, M. Zelechowski, G. Shaikovski, N. Tenenholtz, J. Hall, D. Klimstra, R. Yousfi, et al., Virchow2: Scaling self-supervised mixed magnification models in pathology, arXiv preprint arXiv:2408.00738 (2024).

[22] Bioptimus, H-optimus-1 (2025).
URL https://huggingface.co/bioptimus/H-optimus-1

[23] R. J. Chen, T. Ding, M. Y. Lu, D. F. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, et al., Towards a general-purpose foundation model for computational pathology, Nature medicine 30 (3) (2024) 850–862.

[24] J. Ma, Z. Guo, F. Zhou, Y. Wang, Y. Xu, J. Li, F. Yan, Y. Cai, Z. Zhu, C. Jin, et al., Towards a generalizable pathology foundation model via unified knowledge distillation, arXiv preprint arXiv:2407.18449 (2024).

[25] M. Ranzinger, G. Heinrich, J. Kautz, P. Molchanov, Am-radio: Agglomerative vision foundation model reduce all domains into one, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 12490–12500.

[26] M. B. Sarıyıldız, P. Weinzaepfel, T. Lucas, D. Larlus, Y. Kalantidis, Unic: Universal classification models via multi-teacher distillation, in: European Conference on Computer Vision, Springer, 2024, pp. 353–371.

[27] M. B. Sarıyıldız, P. Weinzaepfel, T. Lucas, P. de Jorge, D. Larlus, Y. Kalantidis, Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 30084–30094.

[28] Y. Guo, J. Stein, G. Wu, A. Krishnamurthy, Sau-net: A universal deep network for cell counting, in: Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics, 2019, pp. 299–306.

[29] Y. Ma, V. Sanchez, T. Guha, Clip-ebc: Clip can count accurately through enhanced blockwise classification, arXiv preprint arXiv:2403.09281 (2024).

[30] R. Jiang, L. Liu, C. Chen, Clip-count: Towards text-guided zero-shot object counting, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 4535–4545.

[31] D. Liang, J. Xie, Z. Zou, X. Ye, W. Xu, X. Bai, Crowdclip: Unsupervised crowd counting via vision-language model, in: Proceedings of

the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 2893–2903.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[33] Z. Huang, Y. Ding, G. Song, L. Wang, R. Geng, H. He, S. Du, X. Liu, Y. Tian, Y. Liang, et al., Bcdata: A large-scale dataset and benchmark for cell detection and counting, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2020, pp. 289–298.

[34] F. Yan, Q. Da, H. Yi, S. Deng, L. Zhu, M. Zhou, Y. Liu, M. Feng, J. Wang, X. Wang, et al., Artificial intelligence-based assessment of pd-l1 expression in diffuse large b cell lymphoma, NPJ Precision Oncology 8 (1) (2024) 76.

[35] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, A. Joulin, Emerging properties in self-supervised vision transformers, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9650–9660.

[36] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, T. Kong, ibot: Image bert pre-training with online tokenizer, arXiv preprint arXiv:2111.07832 (2021).

[37] S. Azizi, L. Culp, J. Freyberg, B. Mustafa, S. Baur, S. Kornblith, T. Chen, N. Tomasev, J. Mitrović, P. Strachan, et al., Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging, Nature Biomedical Engineering 7 (6) (2023) 756–779.

[38] E. Vorontsov, A. Bozkurt, A. Casson, G. Shaikovski, M. Zelechowski, K. Severson, E. Zimmermann, J. Hall, N. Tenenholtz, N. Fusi, et al., A foundation model for clinical-grade computational pathology and rare cancers detection, Nature medicine 30 (10) (2024) 2924–2935.

[39] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu, et al., A whole-slide foundation model for

digital pathology from real-world data, Nature 630 (8015) (2024) 181–188.

[40] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[41] A. Filiot, N. Dop, O. Tchita, A. Riou, R. Dubois, T. Peeters, D. Valter, M. Scalbert, C. Saillard, G. Robin, et al., Distilling foundation models for robust and efficient models in digital pathology, arXiv preprint arXiv:2501.16239 (2025).

[42] C. Zeng, Y. Jiang, F. Zhang, A. Gambaruto, T. Burghardt, Agglomerating large vision encoders via distillation for vfss segmentation, arXiv preprint arXiv:2504.02351 (2025).

[43] G. Heinrich, M. Ranzinger, H. Yin, Y. Lu, J. Kautz, A. Tao, B. Catanzaro, P. Molchanov, Radiov2. 5: Improved baselines for agglomerative vision foundation models, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 22487–22497.

[44] M. Bilal, R. Jewsbury, R. Wang, H. M. AlGhamdi, A. Asif, M. Eastwood, N. Rajpoot, An aggregation of aggregation methods in computational pathology, Medical image analysis 88 (2023) 102885.

[45] L. Hou, D. Samaras, T. M. Kurc, Y. Gao, J. E. Davis, J. H. Saltz, Patch-based convolutional neural network for whole slide tissue image classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2424–2433.

[46] X. Luo, X. Wang, F. Eweje, X. Zhang, S. Yang, R. Quinton, J. Xiang, Y. Li, Y. Ji, Z. Li, et al., Ensemble learning of foundation models for precision oncology, arXiv preprint arXiv:2508.16085 (2025).

[47] L. Solorzano, S. Robertson, B. Acs, J. Hartman, M. Rantalainen, Ensemble-based deep learning improves detection of invasive breast cancer in routine histopathology images, Heliyon 10 (12) (2024).

[48] R. J. Chen, M. Y. Lu, J. Wang, D. F. Williamson, S. J. Rodig, N. I. Lindeman, F. Mahmood, Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and

prognosis, IEEE Transactions on Medical Imaging 41 (4) (2020) 757–770.

[49] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.

[50] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, arXiv preprint arXiv:2304.07193 (2023).

[51] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[52] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[53] B. Wang, H. Liu, D. Samaras, M. H. Nguyen, Distribution matching for crowd counting, Advances in neural information processing systems 33 (2020) 1595–1607.

[54] L. Liu, H. Lu, H. Xiong, K. Xian, Z. Cao, C. Shen, Counting objects by blockwise classification, IEEE Transactions on Circuits and Systems for Video Technology 30 (10) (2019) 3513–3527.

[55] B. Wang, H. Liu, D. Samaras, M. H. Nguyen, Distribution matching for crowd counting, Advances in neural information processing systems 33 (2020) 1595–1607.

[56] H. Naji, L. Sancere, A. Simon, R. Büttner, M.-L. Eich, P. Lohneis, K. Bożek, Holy-net: Segmentation of histological images of diffuse large b-cell lymphoma, Computers in Biology and Medicine 170 (2024) 107978.

[57] J. Gao, L. Zhao, X. Li, Nwpu-moc: A benchmark for fine-grained multicategory object counting in aerial images, IEEE Transactions on Geoscience and Remote Sensing 62 (2024) 1–14.

[58] Y. Li, X. Zhang, D. Chen, Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1091–1100.

[59] W. Li, X. Huang, Z. Zhu, Y. Tang, X. Li, J. Zhou, J. Lu, Ordinalclip: Learning rank prompts for language-guided ordinal regression, Advances in Neural Information Processing Systems 35 (2022) 35313–35325.

[60] Q. D. Vu, S. Graham, T. Kurc, M. N. N. To, M. Shaban, T. Qaiser, N. A. Koohbanani, S. A. Khurram, J. Kalpathy-Cramer, T. Zhao, et al., Methods for segmentation and classification of digital microscopy tissue images, Frontiers in bioengineering and biotechnology 7 (2019) 53.

[61] N. Kumar, R. Verma, D. Anand, Y. Zhou, O. F. Onder, E. Tsougenis, H. Chen, P.-A. Heng, J. Li, Z. Hu, et al., A multi-organ nucleus segmentation challenge, IEEE transactions on medical imaging 39 (5) (2019) 1380–1391.

[62] J. Gamper, N. Alemi Koohbanani, K. Benet, A. Khuram, N. Rajpoot, Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification, in: European congress on digital pathology, Springer, 2019, pp. 11–19.

[63] A. Mahbod, C. Polak, K. Feldmann, R. Khan, K. Gelles, G. Dorffner, R. Woitek, S. Hatamikia, I. Ellinger, Nuinsseg: a fully annotated dataset for nuclei instance segmentation in h&e-stained histological images, Scientific Data 11 (1) (2024) 295.

[64] M. Schuiveling, H. Liu, D. Eek, G. E. Breimer, K. P. Suijkerbuijk, W. A. Blokx, M. Veta, A novel dataset for nuclei and tissue segmentation in melanoma with baseline nuclei segmentation and tissue segmentation benchmarks, GigaScience 14 (2025) giaf011.

[65] P. Naylor, M. Laé, F. Reyal, T. Walter, Segmentation of nuclei in histopathology images by deep regression of the distance map, IEEE transactions on medical imaging 38 (2) (2018) 448–459.