WHY PATCH SELECTION RARELY SAVES COMPUTE, AND HOW TO FIX IT

Anonymous authors
Paper under double-blind review

ABSTRACT

GigaPixel Whole slide image (WSI) analysis is computationally expensive because pathology foundation models have to process large numbers of image patches extracted from WSI. Recent advances tend to select only a candidate patches for slide-level aggregation and representation. However, the selection process is usually carried on high-magnification resolution which requires heavy computation for the task processing. To avoid such computation burden, the process of patch selection/embedding can be utilized at low-magnification level of WSI; but coming with the price of losing pixel information. We address this problem with a three-stage coarse-to-fine cascaded process to progressively select patch regions from low- to high- magnification levels in WSI; imitating a very similar process of pathologists during microscopy diagnosis. A frozen thumbnail encoder first screens all candidate regions (i.e. global tissue view), followed by a frozen pathology foundation model that re-ranks the survivors using one coarse view per region. Only the final selected regions are then encoded at higher resolution. All encoders are frozen, and two lightweight linear attention scorers are trained with attention distillation to transfer the coarse ranking to the thumbnail stage. Across three classification and five survival datasets, we demonstrate the our cascade processing significantly reduces the inference GPU time and resource by ~ 14–$28\times$, while preserving the task performance evaluation.

1 INTRODUCTION

Whole slide images contain gigapixels of tissue and are typically analyzed by dividing each slide into thousands of image patches. Pathology foundation models such as UNI-v2 and Virchow-v2 (Chen et al., 2024; Zimmermann et al., 2024) provide strong representations for these patches, but applying them to every candidate patch can make WSI inference expensive. A natural response is to keep only a small number of informative patches. The problem is that selecting fewer patches does not necessarily mean that fewer patches need to be encoded. The cost depends on how the selection is performed. Suppose a slide has N candidate regions and only $b \ll N$ regions are eventually retained. If the selector uses fine-resolution features, all N regions must first pass through the expensive pathology foundation model. Although the final bag contains only b regions, the foundation model has still been evaluated N times. Selection only reduces the expensive computation when rejected regions can be discarded before fine-resolution encoding. This makes the representation used for selection as important as the final patch budget. (Fig. 1)

We build on this idea with a three-stage coarse-to-fine cascade. First, a frozen DINOv2 thumbnail encoder provides a cheap representation of the entire slide and screens all N candidate regions, retaining only K regions. Second, a frozen pathology foundation model encodes the K survivors once at a coarse resolution and reranks them, retaining only b regions. Finally, the b selected regions are encoded at fine resolution for downstream prediction. The resulting hierarchy is $N \xrightarrow{\text{thumbnail}} K \xrightarrow{\text{coarse resolution}} b \xrightarrow{\text{fine resolution}} \text{prediction}, \quad b \ll K \ll N$. All encoders remain frozen, while two lightweight linear attention scorers provide the selection functions. We further use attention distillation to transfer the coarse pathology ranking to the thumbnail tier, allowing the inexpensive representation to approximate the ranking produced by the more informative coarse representation. Beyond the cascade itself, we use this setup to examine an important question about patch selection: does a smaller bag help because the selected patches are better, or because fewer

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

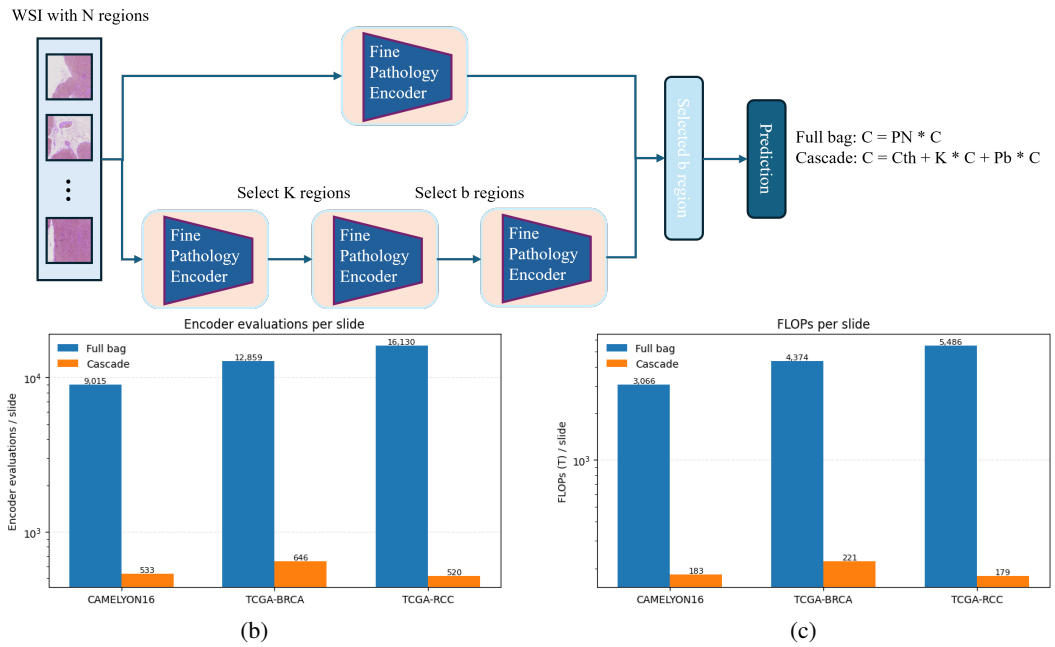

Figure 1: (a) Motivation: reducing the final bag size does not reduce computation when all candidate regions must first be encoded at the fine resolution for selection. (b) Number of pathology encoder evaluations per slide for full-bag and cascade inference across three WSI cohorts. (c) Corresponding encoder FLOPs per slide. The cascade avoids fine-resolution encoding of most candidate regions by progressively filtering them using cheaper representations.

patches need to be processed at fine resolution? To separate these effects, we compare thumbnail, coarse, fine-resolution, random, and oracle selection under the same final patch budgets. We evaluate both the quality of the resulting MIL bag and the computation required to obtain it. Further, we evaluate the cascade on three classification datasets and five survival datasets, using multiple MIL aggregators and selection budgets. The cascade achieves performance close to full-bag inference while substantially reducing the number of fine-resolution foundation-model evaluations and the associated FLOPs. The results also show that the final patch budget alone is not a sufficient measure of inference cost. A small bag can still be expensive if the selection procedure requires encoding the entire slide at the target resolution.

Our contributions are as follows.

- We distinguish the selection budget from the fine-resolution encoding budget and show why retaining fewer patches does not necessarily reduce foundation-model computation.
- We develop a three-stage coarse-to-fine selection cascade that uses a thumbnail representation, a coarse pathology representation, and finally fine-resolution encoding only for the selected regions.
- We design a controlled comparison of full-bag, fine-resolution, coarse, thumbnail, random, and oracle selection to separate subset size from ranking quality.
- We evaluate computational cost and downstream performance across classification and survival tasks, multiple WSI datasets, MIL aggregators, and selection budgets.

2 Related Work

MIL and Patch Selection for WSI. Multiple instance learning (MIL) is widely used for WSI analysis by treating image patches as instances within a slide-level bag. Attention-based MIL learns instance weights for aggregation (Ilse et al., 2018), while later methods introduce clustering and transformer-based aggregation (Lu et al., 2021; Shao et al., 2021). These approaches typically op-

erate on patch representations that have already been extracted. Patch selection methods such as CLAM (Lu et al., 2021) further identify informative regions using learned scores or attention. We study the cost of obtaining the patch representations used for selection. In particular, selecting a small final bag does not reduce fine-resolution encoding cost when all candidate regions must first be processed by the expensive encoder.

Coarse to Fine and Multiscale WSI Inference Several approaches reduce WSI computation by exploiting information across spatial resolutions (Katharopoulos & Fleuret, 2019; Chen et al., 2022; Buzzard et al., 2024). ZoomMIL (Thandiackal et al., 2022) uses hierarchical zooming with gated attention and differentiable Top-K to progressively select regions for higher magnification. Other methods use low-magnification or compressed representations to guide high-resolution processing (Li et al., 2023). Most closely related, WISE-FUSE (Shin et al., 2025) performs coarse-to-fine selection using vision–language similarity between low-resolution patches and class-specific text prompts, then fuses selected patches with textual features. In contrast, our language-free cascade uses a frozen thumbnail for initial screening, followed by coarse pathology features and fine-resolution encoding. We further make the encoder-evaluation cost explicit and compare selection at different representation levels under matched budgets and the same downstream predictor.

Pathology Foundation Models Pathology foundation models such as UNI2 and Virchow2 provide transferable representations across diverse pathology tasks (Chen et al., 2024; Zimmermann et al., 2024). Their increasing representational capacity also makes exhaustive WSI encoding computationally expensive. We use a frozen pathology foundation model as the fine resolution representation and investigate how cheaper representations can be used to avoid unnecessary fine resolution encoding.

3 METHODOLOGY

Problem Formulation We divide a WSI into N non-overlapping candidate regions, forming the candidate set $\mathcal{X} = \{x_i\}_{i=1}^{N}$. Each region is represented at a coarse magnification M_c by a single foundation-model patch and, at the fine magnification $M_f > M_c$, by the set of fine-resolution patches covering the same tissue. With magnification ratio $\rho = M_f/M_c$, a region nominally maps to ρ^2 fine-resolution patches, but because we associate fine patches to a region through a coordinate-matching window rather than an exact grid partition (detail in Appendix), the matching neighborhood is up to $(\rho+1)^2$ patches; the number actually encoded, m_i, is the tissue-containing subset, which is smaller and tissue-dependent, and we measure it per slide rather than assume a constant. Given a final region budget $b \ll N$, our goal is to select a subset $\mathcal{S}_b \subseteq \{1, \ldots, N\}$ with $|\mathcal{S}_b| = b$ before performing fine-resolution feature extraction.

Thumbnail Screening We first construct a low-cost slide-level representation from a single WSI thumbnail. A frozen DINOv2 ViT-B/14 encoder with register tokens produces $Z^{\mathrm{th}} = f_{\mathrm{th}}(x^{\mathrm{th}}) \in \mathbb{R}^{M \times D_{\mathrm{th}}}$, where the spatial token grid contains $M = 73 \times 73 = 5329$ tokens. Each candidate region is associated with the thumbnail token or tokens corresponding to its spatial location. Let $\Pi(i)$ denote this deterministic mapping. The region-level thumbnail representation is $z_i^{\mathrm{th}} = \frac{1}{|\Pi(i)|} \sum_{j \in \Pi(i)} Z_j^{\mathrm{th}}$. When a region corresponds to a single spatial token, this reduces to a direct token lookup.

Coarse Representation The thumbnail scorer provides a cheap first-stage ranking, but its representation lacks fine-resolution pathology features. We therefore rerank the surviving regions using a coarse view of the pathology foundation model. We read each candidate region at the coarse magnification M_c as a single patch and encode it with the frozen pathology foundation model f_{PF} to obtain the region-level coarse feature z_i^c. The same foundation-model weights are used at both tiers; only the input magnification differs. A region therefore costs a single foundation-model evaluation at the coarse tier, whereas encoding it at the fine resolution requires one evaluation per fine-resolution patch it contains. The coarse tier is thus cheaper per region by the number of fine-resolution patches per region, m_i (nominally ρ^2, and up to $(\rho+1)^2$ under our matching window).

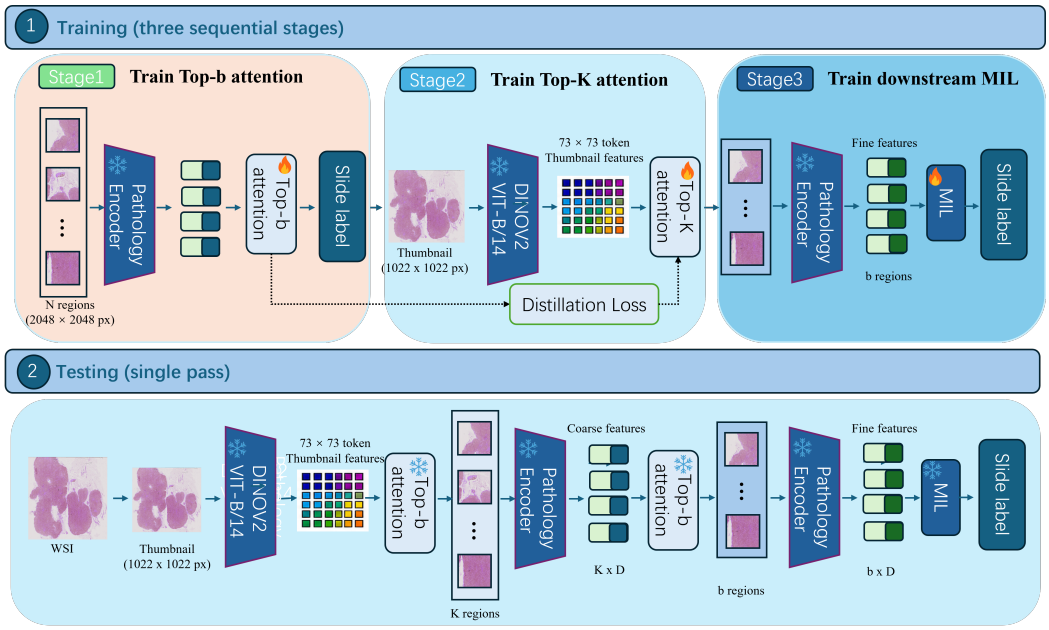

Figure 2: Overview of the coarse-to-fine cascade. A single frozen DINOv2 thumbnail encoder scores all N candidate regions and retains the top K; a frozen pathology foundation model scores a coarse ($5\times$) view of each survivor and retains the top b; only the final b regions are encoded at the fine resolution ($20\times$) by the same foundation model. Coarse scoring is one evaluation per region; fine-resolution encoding is one evaluation per constituent patch. All encoders are frozen; only the two linear attention scorers and the downstream head are trained. The dashed path is the coarse-to-thumbnail attention distillation (Eq. 3), used during training only.

Fine-Resolution Representation For a selected region x_i, we extract its fine-resolution patch embeddings using the same frozen pathology foundation encoder,$e_{ij} = f_{\mathrm{PF}}(p_{ij}), \quad p_{ij} \in \mathcal{P}_f(x_i), \quad j = 1, \ldots, m_i$, where $\mathcal{P}_f(x_i)$ is the set of m_i fine-resolution patches covering the region and m_i is the (tissue-dependent) number actually present. The embeddings are aggregated within the region using $\mathcal{A}_f$ to obtain $z_i^f = \mathcal{A}_f(e_{i1}, \ldots, e_{im_i})$. These expensive representations are computed only after the final region subset has been determined.

Algorithm 1: Coarse-to-Fine WSI Inference

Require: WSI W, budgets K and b
Ensure: Prediction $\hat{y}$
1: Partition W into N regions $\mathcal{R} = \{r_i\}_{i=1}^N$
2: $X^{\mathrm{th}} \leftarrow f_{\mathrm{th}}(\mathrm{Thumb}(W))$
3: Compute thumbnail scores $\{s_i^{\mathrm{th}}\}_{i=1}^N$
4: $\mathcal{I}_K \leftarrow \mathrm{TopK}(\{s_i^{\mathrm{th}}\}_{i=1}^N, K)$
5: $X^c \leftarrow f_c(\{r_i : i \in \mathcal{I}_K\})$
6: $\mathcal{I}_b \leftarrow \mathrm{TopK}(\{s_i^c\}_{i \in \mathcal{I}_K}, b)$
7: Construct fine-resolution patch union $\mathcal{P}_b$
8: $X^t \leftarrow f_t(\mathcal{P}_b)$
9: $\hat{y} \leftarrow \mathrm{MIL}(X^t)$
10: **return** $\hat{y}$

Region Scoring For each representation tier, we use a lightweight linear attention scorer. Given candidate representations $Z = [z_1, \ldots, z_n]^\top$, the scorer produces a scalar score $q_i = v^\top z_i$ and normalized attention $a_i = \mathrm{softmax}(q_i)$. The corresponding bag representation is $h = \sum_i a_i z_i$, followed by a task-specific prediction $\theta = w^\top h + \beta$. For the thumbnail and coarse tiers, the region scores are therefore $q_i^{\mathrm{th}} = (v^{\mathrm{th}})^\top z_i^{\mathrm{th}}, \quad q_i^c = (v^c)^\top z_i^c$. Only the raw scores are required for ranking. The normalized attention weights are used for bag aggregation and attention distillation.

Progressive Region Selection Let $\mathcal{I}_0 = \{1, \ldots, N\}$ denote all candidate regions. The thumbnail scorer first retains the top K candidates, $\mathcal{I}_K = \mathrm{TopK}\left(\{q_i^{\mathrm{th}}\}_{i \in \mathcal{I}_0}, K\right)$. The coarse representation is then computed only for these K candidates and used for a second ranking, $\mathcal{I}_b = \mathrm{TopK}\left(\{q_i^c\}_{i \in \mathcal{I}_K}, b\right)$. Finally, only regions in $\mathcal{I}_b$ are expanded into fine-resolution patches. The final fine-resolution bag is $Z_b^f = \{z_i^f \mid i \in \mathcal{I}_b\}$ and is passed to the downstream MIL predictor. This

ordering is important because the expensive fine-resolution representations are never computed for candidates rejected by either earlier selection stage.

Inference Algorithm 1 summarizes inference. The thumbnail encoder runs once per slide, the coarse encoder runs only on the K survivors, and the fine encoder runs only on the fine-resolution patch union $\mathcal{P}_b$ of the b selected regions. Because each selected region expands into several fine-resolution patches, the fine tier performs $P_b = |\mathcal{P}_b|$ evaluations rather than b, a distinction the cost analysis below makes precise.

Selection Location Determines Encoding Cost The central computational distinction of our method is between the number of regions retained and the number of expensive representations computed before selection. Let c_{enc} denote the cost of one pathology foundation-model evaluation and c_{th} the cost of the thumbnail encoder. Both tiers apply the same foundation model to a fixed 224×224 input, so every foundation-model evaluation costs the same c_{enc} in coarse or fine tier. The tiers differ only in how many evaluations they require. The coarse tier costs one evaluation per scored region; the fine tier costs one evaluation per fine-resolution patch that is actually encoded. We write P_{full} for the total number of fine-resolution patches in a slide and P_b for the number encoded for the selected budget, i.e. the deduplicated union of the fine-resolution patches of the b chosen regions. Both are measured per slide from the extraction logs rather than assumed. The encoder-level costs of full-bag and cascade inference are then

$$C_{\mathrm{full}} = P_{\mathrm{full}}\, c_{\mathrm{enc}}, \qquad C_{\mathrm{cascade}} = c_{\mathrm{th}} + K\, c_{\mathrm{enc}} + P_b\, c_{\mathrm{enc}}. \tag{1}$$

Here the coarse tier contributes K evaluations (one per thumbnail survivor) and the fine tier contributes P_b. Writing $\bar{m}$ for the mean number of fine-resolution patches encoded per selected region, $P_b \approx (1-\eta)\, b\, \bar{m}$, where η is the small overlap between the chosen regions' fine-resolution neighborhoods (a few percent in our data). The value of $\bar{m}$ is tissue-dependent and varies by dataset; we therefore report the measured P_b and P_{full} per slide rather than a fixed multiplier.

The cascade reduces fine-resolution encoding whenever $c_{\mathrm{th}} + K\, c_{\mathrm{enc}} < (P_{\mathrm{full}} - P_b)\, c_{\mathrm{enc}}$. This also explains why the final patch budget alone does not guarantee computational savings. If a selector instead ranks candidates using fine-resolution features, those features must be computed for all N candidates before selection, incurring at least $P_{\mathrm{full}}\, c_{\mathrm{enc}}$; a smaller final bag then reduces the downstream MIL input but not the fine-resolution encoding already spent on selection. The same accounting favors the cascade at training time. Training the coarse scorer needs a coarse view of all N regions, but that is one evaluation per region; only the b selected regions per slide are ever encoded at the fine resolution. The per-slide training encoding cost is therefore

$$C_{\mathrm{train}} = c_{\mathrm{th}} + N\, c_{\mathrm{enc}} + P_b\, c_{\mathrm{enc}}, \tag{2}$$

which stays far below the $P_{\mathrm{full}}\, c_{\mathrm{enc}}$ of encoding every region at the fine resolution, since one coarse evaluation replaces the up-to-ρ^2 fine-resolution evaluations of a region. The cascade thus never performs full fine-resolution encoding, at training or inference; the only difference is that the coarse tier is applied to all N regions during training but only to the K thumbnail survivors at inference, and these coarse and thumbnail features are computed once per slide and cached across epochs.

Equation 1 counts foundation-model evaluations, which dominate encoder compute but are not the only cost. The coarse tier still reads and decodes all N candidate regions from the WSI pyramid, so the cascade avoids their fine-resolution encoding but not their decompression, and region I/O and decompression are themselves a known bottleneck in WSI processing (Li et al., 2023). The evaluation-count savings therefore upper-bound the achievable wall-clock speedup rather than predict it; we report measured wall-clock time separately to account for decompression, batching, and data transfer.

Attention Distillation The coarse scorer is first trained using the complete candidate set and provides an attention distribution a^c. We then train the thumbnail scorer to reproduce this distribution while optimizing the downstream task objective. The distillation loss is

$$\mathcal{L}_{\mathrm{dist}} = -\sum_{i=1}^{N} a_i^c \log a_i^{\mathrm{th}}, \qquad \mathcal{L}_{\mathrm{th}} = \mathcal{L}_{\mathrm{task}} + \lambda \mathcal{L}_{\mathrm{dist}}. \tag{3}$$

This transfers the coarse selection preference to the substantially cheaper thumbnail representation. During inference, however, the coarse scorer is evaluated only on the K thumbnail survivors. All image encoders remain frozen, and only the lightweight scoring and prediction modules are trained.

Task Objectives For binary classification, the final score is converted to a probability $p = \sigma(\theta)$ and optimized using binary cross-entropy, $\mathcal{L}_{\text{cls}} = -y \log p - (1-y)\log(1-p)$. For survival prediction, we use a patient-level risk score θ_i^{pat}. When each patient contributes one slide, this is simply the corresponding slide-level score. With event indicator δ_i and risk set $\mathcal{R}_i$, the Cox objective used to train the selection scorer is

$$\mathcal{L}_{\text{surv}} = -\sum_i \delta_i \left[\theta_i^{\text{pat}} - \log \sum_{j \in \mathcal{R}_i} \exp(\theta_j^{\text{pat}}) \right]. \tag{4}$$

The Cox objective provides a relative patient-level risk ranking for selection. The downstream survival predictor retains the discrete-time survival objective used in the final evaluation.

Downstream Prediction and Evaluation After selection, all methods use the same downstream MIL architecture. Given the selected fine-resolution representations Z_b^f, the final prediction is $\hat{y} = \mathcal{A}_{\text{MIL}}(Z_b^f)$. We evaluate multiple MIL aggregators while keeping the selection budget fixed: mean and max pooling, ABMIL (Ilse et al., 2018), CLAM (Lu et al., 2021), DSMIL (Li et al., 2021), TransMIL (Shao et al., 2021), DTFD-MIL (Zhang et al., 2022), and MambaMIL (Yang et al., 2024). selection budget fixed.

To separate ranking quality from subset cardinality, we compare full-bag encoding, fine-resolution selection, coarse selection, thumbnail selection, the proposed cascade, random selection, and a Tumor Oracle. All selection methods use the same final budget b when ranking quality is compared. The Tumor Oracle ranks regions by ground-truth tumor overlap and therefore applies only to datasets with lesion annotations (CAMELYON16 in our study).

Sensitivity experiments vary K and b independently to distinguish the computational operating point from the general behavior of the selection hierarchy.

All data splits are performed at the patient level, ensuring that slides from the same patient remain within the same partition. Test-slide annotations are never used by the learned selectors.

4 Experiments

Datasets. We evaluate on three classification datasets and five survival datasets. The classification datasets are CAMELYON16 (Bejnordi et al., 2017) (breast lymph-node metastasis detection), TCGA-BRCA (breast carcinoma subtyping), and TCGA-RCC (renal cell carcinoma subtyping). The survival datasets are TCGA-KIRC, TCGA-KIRP, TCGA-LUAD, TCGA-STAD, and TCGA-UCEC, each evaluated by cross-validated concordance index.

Implementation. The frozen patch encoder is UNI-v2 and Virchow2 for classification and survival. The thumbnail encoder is the public DINOv2 ViT-B/14 on a 1022×1022 thumbnail, giving a 73×73 (5,329-token) grid. The cascade uses $K{=}500$ and $b{=}10$, and every method in a comparison is evaluated under $3{\times}5$ split-seed cross-validation in survival and $5{\times}5$ split-seed cross-validation in classification and timed on the same fixed slide subset and GPU. Full details of tiling and feature extraction , the selection network, the training procedure, and hyperparameters are deferred to the appendix.

Survival. Table 1 reports the concordance index across the five survival datasets for full-bag encoding and the cascade. The cascade tracks full-bag encoding closely. We treat survival as evidence that the cascade preserves downstream utility on a different task family under a much smaller encoding budget. Results with a second encoder (Virchow2) and wall clock result are in Appendix.

Classification. Table 2 and Figure 3 reports AUROC and GPU-forward time for full-bag encoding, four published selection baselines (ZoomMIL (Thandiackal et al., 2022), HAG-MIL (Xiong et al.,

Table 1: Survival: concordance index (C-index) across five TCGA datasets using UNI-v2, mean ± fold-level standard error over 15 runs (3 splits × 5 seeds). Full bag encodes every region; Cascade uses $K = 500$, $b = 10$; all arms share the downstream loss, model-selection rule, and splits.

Method	TCGA-KIRC	TCGA-KIRP	TCGA-LUAD	TCGA-STAD	TCGA-UCEC	Avg.
Mean	$0.751_{\pm 0.032}$	$0.816_{\pm 0.006}$	$0.595_{\pm 0.036}$	$0.654_{\pm 0.033}$	$0.687_{\pm 0.136}$	0.701
Max	$0.708_{\pm 0.044}$	$0.578_{\pm 0.020}$	$0.548_{\pm 0.045}$	$0.548_{\pm 0.050}$	$0.699_{\pm 0.112}$	0.616
ABMIL	$0.741_{\pm 0.039}$	$0.821_{\pm 0.018}$	$0.601_{\pm 0.039}$	$0.641_{\pm 0.035}$	$0.635_{\pm 0.110}$	0.688
TransMIL	$0.733_{\pm 0.031}$	$0.807_{\pm 0.021}$	$0.573_{\pm 0.019}$	$0.625_{\pm 0.019}$	$0.706_{\pm 0.112}$	0.689
CLAM	$0.744_{\pm 0.035}$	$0.821_{\pm 0.025}$	$0.599_{\pm 0.033}$	$0.644_{\pm 0.022}$	$0.665_{\pm 0.137}$	0.695
Cascade(Mean)	$0.745_{\pm 0.045}$	$0.795_{\pm 0.017}$	$0.576_{\pm 0.033}$	$0.604_{\pm 0.017}$	$0.728_{\pm 0.033}$	0.690
Cascade(Max)	$0.702_{\pm 0.042}$	$0.651_{\pm 0.045}$	$0.562_{\pm 0.041}$	$0.558_{\pm 0.023}$	$0.700_{\pm 0.042}$	0.635
Cascade(ABMIL)	$0.743_{\pm 0.047}$	$0.812_{\pm 0.025}$	$0.580_{\pm 0.032}$	$0.599_{\pm 0.020}$	$0.684_{\pm 0.015}$	0.684
Cascade(TransMIL)	$0.722_{\pm 0.043}$	$0.780_{\pm 0.025}$	$0.552_{\pm 0.016}$	$0.578_{\pm 0.013}$	$0.734_{\pm 0.036}$	0.673
Cascade(CLAM)	$0.735_{\pm 0.046}$	$0.808_{\pm 0.022}$	$0.578_{\pm 0.033}$	$0.599_{\pm 0.012}$	$0.732_{\pm 0.031}$	0.690

GPU forward time (full → cascade, s): 509→27 (KIRC), 477→27 (KIRP), 419→22 (LUAD), 417→24 (STAD), and 824→29 (UCEC), corresponding to approximately 17–28× reductions.

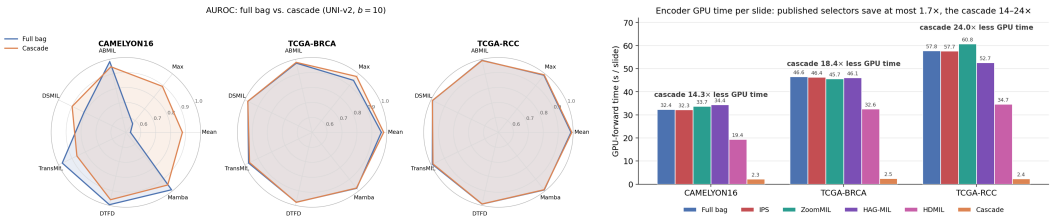

Figure 3: Left: Full-bag vs. cascade AUROC (UNI-v2, b=10) across seven MIL aggregators. On the diffuse subtyping cohorts (TCGA-BRCA, TCGA-RCC) the two polygons nearly coincide, i.e. the cascade matches full-bag encoding; on the detection cohort (CAMELYON16) they diverge, with the cascade improving the weak-pooling heads (Mean, Max, DSMIL) and the full bag retaining a small edge on the attention and transformer heads.Right: Measured encoder forward time per slide using CUDA events, comparing full-bag and cascade inference with UNI-v2, averaged over 10 slides. Speed-ups (14.3/18.4/24.0×) are annotated.

2023), HDMIL (Dong et al., 2025), and IPS (Bergner et al., 2023)), and the cascade across seven MIL aggregators. The cascade matches full-bag encoding on the diffuse subtyping datasets (TCGA-BRCA, TCGA-RCC), trails it on the strongest heads for CAMELYON16 detection, and improves the weak pooling heads (Mean, Max, DSMIL). Results with a second encoder (Virchow2), wall clock and a second coarse level (10×) are in Appendix.

Table 2: Classification AUROC and inference GPU forward time (UNI-v2) across datasets. Time is the time per slide (s) spent inside the encoder forward passes (CUDA-event time, excluding tile reads, decoding and preprocessing), measured on the same 10 benchmark slides per dataset and one GPU. AUROC is reported as mean with fold-level standard error over 25 runs (5 folds × 5 seeds). Cascade results use $K = 500$ and $b = 10$.

	CAM16		TCGA-BRCA		TCGA-RCC	
Method	AUROC	Time	AUROC	Time	AUROC	Time
ZoomMIL	$0.9336_{\pm 0.0042}$	33.7 s	$0.9808_{\pm 0.0050}$	45.7 s	$0.9930_{\pm 0.0031}$	60.8 s
HAGMIL	$0.8298_{\pm 0.0355}$	34.4 s	$0.9669_{\pm 0.0084}$	46.1 s	$0.9919_{\pm 0.0036}$	52.7 s
HDMIL	$0.9718_{\pm 0.0060}$	19.4 s	$0.9740_{\pm 0.0055}$	32.6 s	$0.9940_{\pm 0.0031}$	34.7 s
IPS	$0.9767_{\pm 0.0020}$	32.3 s	$0.9795_{\pm 0.0049}$	46.4 s	$0.9868_{\pm 0.0052}$	57.7 s
Cascade (DTFD-MIL)	$0.9613_{\pm 0.0066}$	2.3 s	$0.9792_{\pm 0.0055}$	2.5 s	$0.9903_{\pm 0.0033}$	2.4 s

Table 3: Selection rule at a matched budget (CAMELYON16, UNI-v2 $20\times$ features, b=10, ABMIL, test AUROC over 25 runs). Encoder evals is each rule's pre-selection cost (Eq. 1).

Selection strategy	Encoder evals	AUROC
Full bag (reference)	P_{full}	0.9844
Fine-resolution	P_{full}	0.906
Random	P_b	0.629
Tumour oracle	P_b	0.6423
Thumbnail	$c_{\text{th}}+P_b$	0.6613
Coarse	$N+P_b$	0.9396
Cascade	$c_{\text{th}}+K+P_b$	**0.9491**

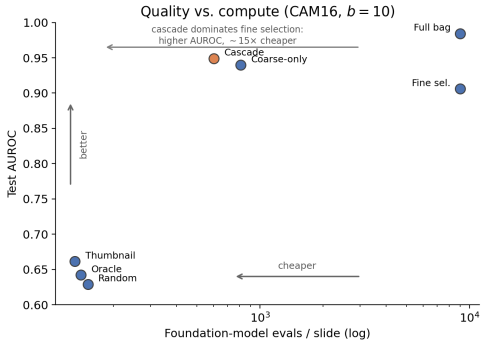

Figure 4: Accuracy vs. encoder cost at b=10 (CAMELYON16). The cascade beats fine-resolution selection with $\sim$15$\times$ fewer evaluations.

4.1 Ablations.

Where should the selection decision be made? Putting all selection strategies under the same patch budget makes the difference clear. Table 3 compares the full-bag reference with fine-resolution, random, tumour-oracle, thumbnail, coarse, and cascade selection at b=10 on CAMELYON16. All methods use the same UNI-v2 $20\times$ features and ABMIL head, so the only difference is how the b regions are selected. The representation-based selectors, coarse and cascade, perform substantially better than random selection. This shows that, even with the same final number of regions, the representation used for selection can have a large effect on accuracy, independently of the patch budget itself. The effect of changing the budget is examined in Appendix.

The tumour oracle provides another useful comparison. It provides only a modest improvement over random selection in this experiment. This suggests that lesion localization alone does not necessarily identify the regions most useful for the downstream task. The fine-resolution baseline highlights the computational trade-off. It encodes all N regions at $20\times$ before ranking them, paying essentially the full-bag encoding cost, yet still performs below the cascade, with 0.906 compared with 0.949. This is not a hypothetical baseline. IPS is a published selector that follows the same strategy, encoding every $20\times$ patch before selecting a smaller bag. Figure 4 summarizes this trade-off by plotting each selection strategy according to its encoder cost and accuracy.

Two further analyses support this reading (Appendix). Selection quality rises steeply with the coarse-tier magnification from $1\times$ to $5\times$ and plateaus by $10\times$, so the $5\times$ tier sits at the knee and most of the benefit of the pathology encoder is available despite its lower input magnification. On CAMELYON16's pixel annotations, all learned region selectors enrich tumour roughly $40\times$ over chance, yet the fine-resolution selector captures the most tumour while giving the lowest AUROC, so downstream accuracy is decoupled from tumour content at the top of the ranking.

Sensitivity to the thumbnail budget K. The thumbnail budget K determines the cascade's compute operating point, entering the cost model directly through Kc_{enc} (Eq. 1). Its effect on downstream performance is task-dependent (Table 4). On CAMELYON16, increasing K from 100 to 500 substantially improves AUROC, from 0.9044 to 0.9629, while increasing K further to 1000 produces a nearly unchanged result. In contrast, performance on TCGA-BRCA is relatively stable across the tested budgets, with AUROC ranging from 0.9412 to 0.9470, and TCGA-RCC is similarly stable, ranging from 0.9941 to 0.9945. Thus, the benefit of a larger thumbnail candidate pool is most pronounced for the focal detection task, whereas dif-

Table 4: Thumbnail-budget sensitivity at b=10 (ABMIL, val AUROC, 25 runs). Performance peaks near K=500 on focal detection (CAMELYON16) and is flat in K on diffuse subtyping.

Dataset	K=100	K=250	K=500	K=1000
CAMELYON16	0.9044	0.9445	**0.9629**	0.9618
TCGA-BRCA	**0.9470**	0.9452	0.9423	0.9412
TCGA-RCC	0.9943	**0.9945**	**0.9945**	0.9941

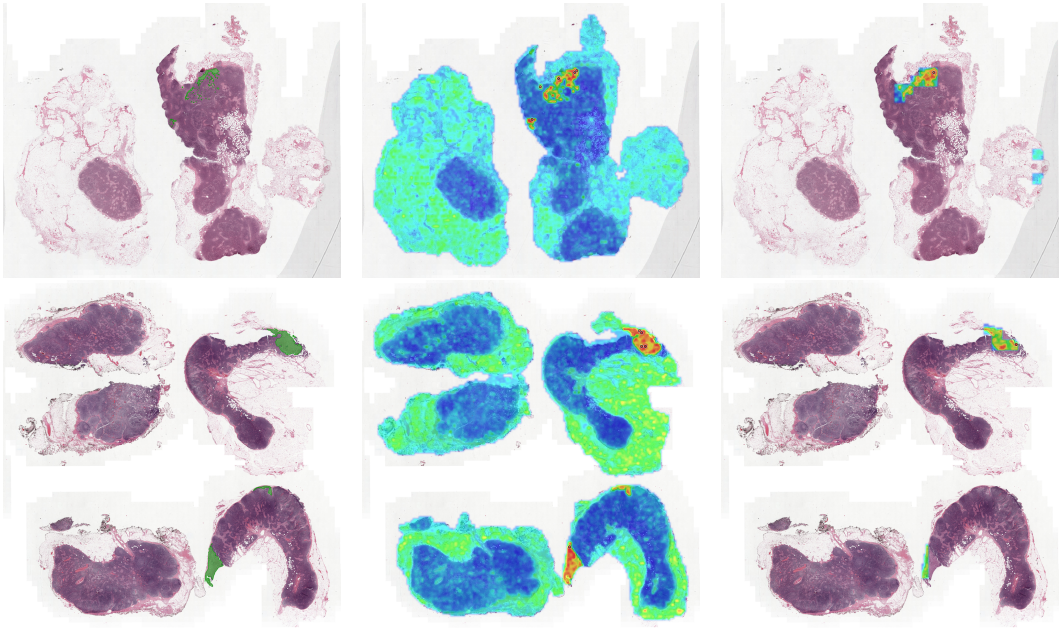

Figure 5: Attention visualizations of ABMIL and Cascade(ABMIL) on two CAMELYON16 samples for classification. Each row shows the original WSI region with tumor annotation, full-bag ABMIL attention, and Cascade(ABMIL) attention, respectively. Tumor regions are outlined in green in the first panel. In the second and third panels, blue indicates low attention (less important patches), while red indicates high attention (critical patches).

fuse subtyping is comparatively insensitive to K. Because the coarse tier performs approximately $\min(N, K)$ foundation-model evaluations per slide, increasing K also increases selection cost. We therefore use K=500 as a practical operating point that provides strong performance on CAMELYON16 while remaining within the stable performance range observed on the subtyping datasets. Distillation is necessary but its strength is not a sensitive knob (0.812 at λ=0 vs. 0.943–0.950 for $\lambda \in [0.25, 2]$).

Qualitative Evaluation. We qualitatively compare the attention maps produced by full-bag ABMIL and Cascade(ABMIL) on two CAMELYON16 samples. Figure 5 shows the original regions with tumor annotations alongside the attention maps from the two models. We find that high-attention regions from the cascade overlap visible tumor annotations, while suppressing many non-tumor regions that receive higher attention under full-bag encoding. This indicates that the cascade can preserve task-relevant spatial focus despite evaluating substantially fewer target-resolution regions. The comparison also illustrates that the cascade does not simply reproduce the full-bag attention pattern, but can produce a more concentrated selection of regions for downstream prediction.

5 Discussion and Conclusion

Our results show that WSI selection cost depends on the representation used for selection, not only on the final patch budget. The proposed cascade progressively filters regions before expensive fine-resolution encoding, substantially reducing foundation-model evaluations while remaining competitive with full-bag inference. Matched-budget experiments show that coarse pathology features provide better selection than random or thumbnail-only ranking. Sensitivity to K highlights the trade-off between computational cost and early rejection of informative regions. However, performance can decline on focal detection with overly small K, and encoder-level timing excludes WSI I/O and preprocessing costs. The end to end wall clock time reported in Appendix. Overall, efficient WSI inference should consider not only how many regions are retained, but also the representation and computation required to select them.

AI USE STATEMENT

We used generative AI tools (large language model assistants) for writing assistance, including language editing and restructuring of the manuscript text, for help with LaTeX formatting, for drafting the plotting scripts that generate the figures from our logged results, and for checking bibliographic entries. We have not used generative AI tools to generate experimental results; all reported numbers come from our own experiment logs. We have reviewed all AI-assisted work: every number in the paper was checked against our experiment logs, the plotting code was run and its output inspected by the authors, and every bibliographic entry was checked against the original publication. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available, de-identified histopathology datasets (CAMELYON16 and TCGA) under their respective data-use terms; no new human subjects data were collected and no additional ethics approval was required. The models studied are research prototypes for slide-level classification and survival prediction and are not validated for clinical use. Because patch selection discards most of the tissue, errors on underrepresented tissue types or acquisition sites could be amplified, and any clinical deployment would require prospective validation across institutions and patient populations.

REPRODUCIBILITY STATEMENT

The cascade is specified in Section 3, with the full inference procedure in Algorithm 1 and the cost model in Eq. 1. Section 4 lists the datasets, frozen encoders, budgets (K=500, b=10) and cross-validation protocol (5×5 split-seed runs for classification, 3×5 for survival). Tiling, feature extraction, training hyperparameters and the timing protocol are described in the appendix. All datasets are public, and the thumbnail and pathology encoders are publicly released pretrained models.

REFERENCES

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017.

Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Iterative patch selection for high-resolution image recognition. In *International Conference on Learning Representations (ICLR)*, 2023. github.com/benbergner/ips.

Zak Buzzard, Konstantin Hemker, Nikola Simidjievski, and Mateja Jamnik. Paths: A hierarchical transformer for efficient whole slide image analysis. *arXiv preprint arXiv:2411.18225*, 2024.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. 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*, 30:850–862, 2024.

Jiuyang Dong, Junjun Jiang, Kui Jiang, Jiahan Li, and Yongbing Zhang. Fast and accurate gigapixel pathological image classification with hierarchical distillation multi-instance learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 30818–30828, 2025. arXiv:2502.21130.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning (ICML)*, 2018.

Angelos Katharopoulos and François Fleuret. Processing megapixel images with deep attention-sampling models. In *International Conference on Machine Learning (ICML)*, 2019.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. Includes a pyramidal multiscale-fusion mechanism, referred to elsewhere in this paper as "DSMIL-multiscale".

Zheyu Li, Bin Li, Kevin W. Eliceiri, and Vijaykrishnan Narayanan. Computationally efficient adaptive decompression for whole slide image processing. *Biomedical Optics Express*, 14(2):667–686, 2023.

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

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yonghan Shin, SeungKyu Kim, and Won-Ki Jeong. Wise-fuse: Efficient whole slide image encoding via coarse-to-fine patch selection with vlm and llm knowledge fusion. *arXiv preprint arXiv:2508.14537*, 2025.

Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew F. K. Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *European Conference on Computer Vision (ECCV)*, 2022. arXiv:2204.12454.

Conghao Xiong, Hao Chen, Joseph J. Y. Sung, and Irwin King. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1587–1595, 2023.

Shu Yang, Yihui Wang, and Hao Chen. MambaMIL: Enhancing long sequence modeling with sequence reordering in computational pathology. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 15004 of *Lecture Notes in Computer Science*, pp. 296–306, 2024.

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

Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, et al. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.