Act Like a Pathologist: Tissue-Aware Whole Slide Image Reasoning

Supplementary Material

In this supplementary material, we provide additional technical derivations, experimental details, and qualitative analyses to complement the main manuscript. First, Section 6 presents the mathematical derivation of our Hierarchical Information Bottleneck objective. Subsequently, we introduce the details of our in-house ovarian dataset and the specialized evaluation tool developed for pathologist assessment in Section 7 and Section 8, respectively. We then elaborate on the implementation details of our two-stage training strategy in Section 9. To further demonstrate the model's performance, Section 10 provides additional quantitative results on public benchmarks, followed by extensive qualitative visualizations on both public and private datasets in Section 11. A comprehensive ablation study investigating various training configurations and sampling distributions is presented in Section 12. Finally, we discuss the limitations of our current approach and suggest future research directions in Section 13.

6. Hierarchical IB Objective Derivation

6.1. Variational Information Bound

By definition, the mutual information $I(A;B)$ between two random variables A and B is:

$$I(A;B) = \mathbb{E}_{p(A,B)}\left[\log \frac{p(A \mid B)}{p(A)}\right] \tag{13}$$

In practice, the true marginal distribution $p(A)$ is typically intractable. To derive a computable bound, we introduce a variational prior $q(A)$. Utilizing the non-negativity of the KL divergence:

$$D_{\mathrm{KL}}(p(A) \parallel q(A)) = \mathbb{E}_{p(A)}\left[\log \frac{p(A)}{q(A)}\right] \geq 0 \tag{14}$$

it follows that:

$$\mathbb{E}_{p(A)}[\log p(A)] \geq \mathbb{E}_{p(A)}[\log q(A)] \tag{15}$$

Consequently, by substituting this inequality into the definition of $I(A;B)$, we arrive at the variational upper bound for the compression term, following the VIB framework [2]:

$$I(A;B) \leq \mathbb{E}_{p(B)}[D_{\mathrm{KL}}(p(A \mid B) \parallel q(A))] \tag{16}$$

6.2. Hierarchical Variational Decomposition

As stated in the main text, the total compression term $I(Z;X \mid \mathbf{q})$ is decomposed into group-level and patch-level terms using the chain rule for mutual information:

$$I(Z_g, Z_p; X \mid \mathbf{q}) = I(Z_g; X \mid \mathbf{q}) + I(Z_p; X \mid Z_g, \mathbf{q}) \tag{17}$$

By applying the variational bound from Equation (16) to each hierarchical component, we derive the tractable hierarchical constraints for our selection process. For the group-level complexity, we introduce the variational posterior $p_{\phi_g}(Z_g \mid X, \mathbf{q})$ and the group-level prior $p_g(Z_g \mid \mathbf{q})$. The mutual information term is then bounded by:

$$I(Z_g; X \mid \mathbf{q}) \leq \mathbb{E}_{\mathcal{D}}[D_{\mathrm{KL}}(p_{\phi_g}(Z_g \mid X, \mathbf{q}) \parallel p_g)] \tag{18}$$

Similarly, for the patch-level complexity, we introduce the conditional posterior $p_{\phi_p}(Z_p \mid X, Z_g, \mathbf{q})$ and its corresponding conditional prior $p_p(Z_p \mid Z_g, \mathbf{q})$. The mutual information term is then bounded by:

$$\begin{aligned} I(Z_p; X \mid Z_g, \mathbf{q}) \leq & \mathbb{E}_{\mathcal{D}}\Big[\mathbb{E}_{Z_g \sim p_{\phi_g}}[D_{\mathrm{KL}}(p_{\phi_p}(Z_p \mid X, Z_g, \mathbf{q}) \\ & \parallel p_p(Z_p \mid Z_g, \mathbf{q}))]\Big] \end{aligned} \tag{19}$$

6.3. Deriving the Final Objective

For the relevance term $I(Z;Y \mid \mathbf{q})$, which measures the predictive power of the selected features for the answer Y, we utilize the LLM as a variational decoder $p_\theta(Y \mid Z, \mathbf{q})$ to obtain a tractable lower bound:

$$I(Z;Y \mid \mathbf{q}) \geq \mathbb{E}_{\mathcal{D}}\mathbb{E}_{Z \sim p_\phi}[\log p_\theta(Y \mid Z, \mathbf{q})] \tag{20}$$

By substituting the complexity upper bounds and the relevance lower bound into the original IB objective $\mathcal{L}_{\mathrm{IB}}$, we arrive at a tractable training objective. In the standard VIB [2] formulation, a single hyperparameter β typically regulates the entire compression term. To better accommodate the hierarchical nature of WSIs and provide greater flexibility in hyperparameter tuning, we extend this formulation by assigning independent Lagrange multipliers, β_g and β_p, as group-level and patch-level regularizers, respectively. This decoupling allows the model to independently regulate the information bottleneck at each granularity. The resulting HIB objective is formulated as:

$$\begin{aligned} \mathcal{J}_{\mathrm{HIB}} = \mathbb{E}_{\mathcal{D}}\Big[& \mathbb{E}_{Z \sim p_\phi}[\log p_\theta(Y \mid Z, \mathbf{q})] \\ & - \beta_g D_{\mathrm{KL}}(p_{\phi_g}(Z_g \mid X, \mathbf{q}) \parallel p_g) \\ & - \beta_p \mathbb{E}_{Z_g \sim p_{\phi_g}}[D_{\mathrm{KL}}(p_{\phi_p}(Z_p \mid X, Z_g, \mathbf{q}) \parallel p_p)]\Big] \end{aligned} \tag{21}$$

In practice, the nested expectation over Z_g in the patch-level term is empirically estimated through the group-level sampling process. By using the specific sampling rate made by the group sampler during the forward pass, the theoretical objective reduces to the empirical loss function presented in Equation (8) of the main text.

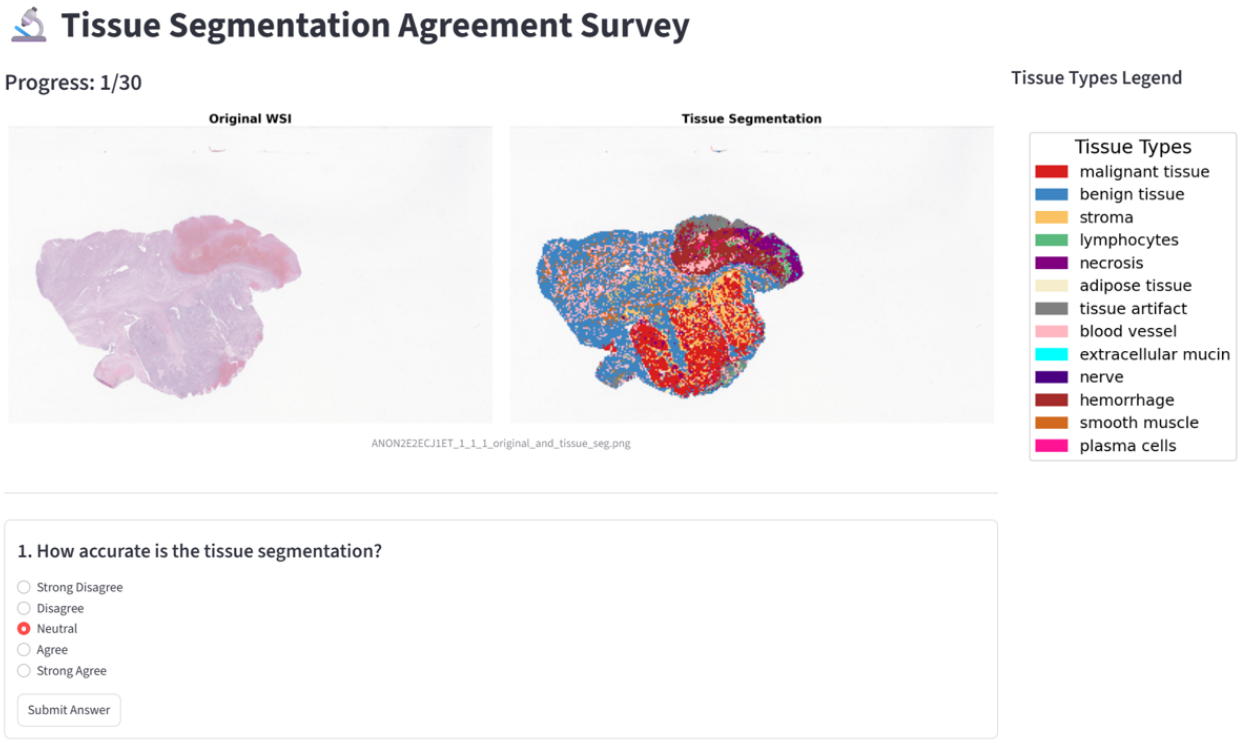

Figure 5. The user interface for the Tissue Segmentation Survey. The central area shows a side-by-side comparison of the original WSI and the tissue segmentation result. The right legend clarifies the tissue classes, and the bottom section collects the pathologists' rating.

7. In-house Ovarian Dataset

To demonstrate the generalizability of our proposed model, we curated a small-scale, in-house ovarian dataset. This dataset is compiled from WSIs of ovarian tissues and formatted into question-answer pairs, focusing on distinct histological phenotypes visible within the WSIs. The dataset includes four primary diagnostic categories, based on the observed tumor morphology. In total, the dataset comprises 375 question-answer pairs. The distribution of samples across the four categories is as follows: endometrioid ($n = 81$), clear cell carcinoma ($n = 82$), high grade serous carcinoma ($n = 123$), and serous borderline carcinoma ($n = 89$).

A typical question within the dataset is structured as a multiple-choice classification task based on visual features observed in the WSI. An example is provided below:

> **Example Question-Answer Pair:** Based on the observed features, what do you think is the correct histological classification of the tumor?
>
> (a) endometrioid
> (b) clear cell carcinoma
> (c) high grade serous carcinoma
> (d) serous borderline carcinoma

8. Evaluation Tool

To verify the reliability of our tissue segmentation and ensure that the model selects question-relevant patches for inference, we developed an interactive survey tool based on Streamlit [1]. This evaluation is twofold: it primarily validates the accuracy of tissue segmentation, followed by an assessment of the patch selection performance. We detail these two components in the following sections.

Tissue Segmentation Verification. In the first part, we focus on the tissue segmentation results. As illustrated in Figure 5, the user interface features a side-by-side comparison view: the original WSI is displayed on the left, while the corresponding model-generated segmentation mask is shown on the right. To assist pathologists in interpreting the results, a color-coded legend is provided on the sidebar, mapping specific colors to tissue types (e.g., red for malignant tissue, yellow for stroma).

Pathologists are asked to verify the alignment between the WSI and the mask using these visual cues. Based on their inspection, they rate the segmentation quality via a radio button selection:

- **Q1:** "How accurate is the tissue segmentation?" (1 = Strongly Disagree, 5 = Strongly Agree)

After the user clicks "Submit Answer", the tool automati-

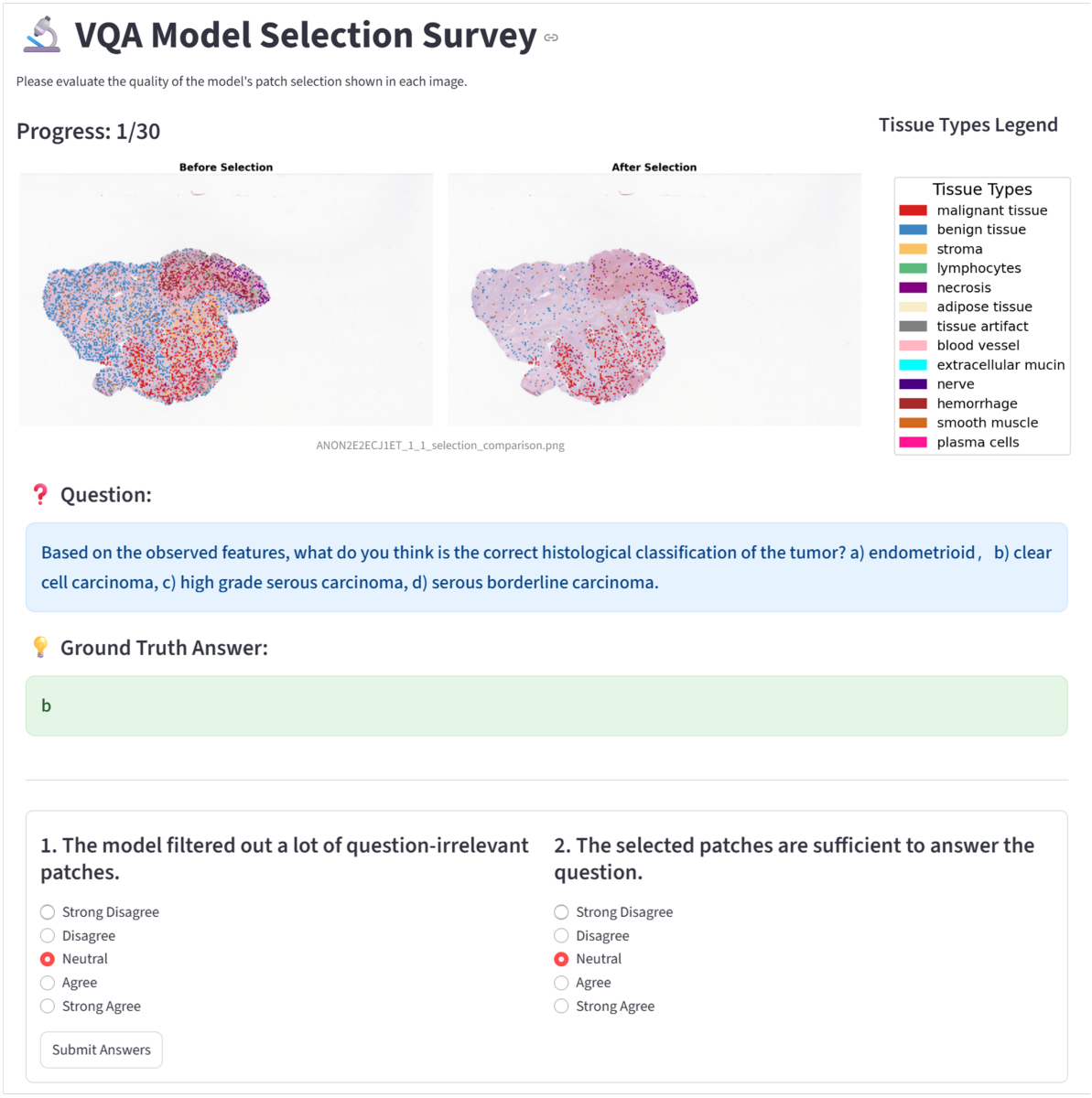

Figure 6. The user interface for the Patch Selection Survey. The view visualizes the "Before" and "After" states of patch selection. Note that the input question and ground truth answer are displayed below the images to provide necessary context for the pathologists' evaluation.

cally logs the feedback and advances to the next sample.

Patch Selection Assessment. In the second part, the tool adapts to evaluate the model's coarse-to-fine selection capability. As shown in Figure 6, we visualize the patches in both their "Before" and "After" selection states.

To ensure sufficient diagnostic context for evaluation, the specific "Question" (highlighted in blue) and the "Ground Truth Answer" (highlighted in green) are explicitly displayed below the images. Pathologists assess whether the model successfully removes irrelevant regions while retaining diagnostic information by answering the following:

- **Q2.1:** "Does the model filter out a significant number of question-irrelevant patches?"
- **Q2.2:** "Are the selected patches sufficient to answer the question?"

To conduct a robust evaluation, we randomly sampled a total of 30 cases, comprising 20 from the public dataset and 10 from the private dataset. Two independent pathologists were invited to perform the annotations using the developed tool. To ensure the reliability and objectivity of the assessment, we calculated the average scores from both annotators as the final metric for pathologist-based evaluation.

Table 7. Quantitative results on SlideBench-VQA (TCGA). Performance is evaluated using accuracy and macro-averaged metrics.

	Microscopy				Diagnosis				Clinical			
Method	Acc.	Macro-P	Macro-R	Macro-F1	Acc.	Macro-P	Macro-R	Macro-F1	Acc.	Macro-P	Macro-R	Macro-F1
GPT-4o	39.24	36.12	39.45	35.54	24.12	22.83	24.34	21.72	44.67	42.93	44.95	43.58
Quilt-LLaVA [6]	52.39	46.75	50.09	46.84	30.19	27.94	30.34	27.15	49.33	47.01	48.91	47.54
LLaVA-Med [4]	52.15	46.30	49.80	46.51	29.97	27.84	29.93	26.86	47.33	45.12	47.19	45.58
SlideChat [3]	83.15	**81.77**	78.50	79.70	71.36	71.80	65.22	67.52	75.33	73.84	72.74	72.98
Ours	**84.62**	80.79	**80.47**	**80.33**	**73.09**	**72.10**	**68.08**	**69.22**	**77.30**	**73.97**	**74.24**	**73.94**

Table 8. Quantitative results on WSI-Bench (Close-ended). Performance is evaluated using accuracy and macro-averaged metrics.

	Morphology				Diagnosis				Treatment (Binary)			
Method	Acc.	Macro-P	Macro-R	Macro-F1	Acc.	Macro-P	Macro-R	Macro-F1	Acc.	Macro-P	Macro-R	Macro-F1
GPT-4o	47.07	42.89	47.29	43.19	53.06	49.02	53.37	49.27	87.50	79.12	83.76	81.05
Quilt-LLaVA [6]	94.13	91.74	91.19	91.42	84.13	81.35	78.68	79.63	**97.92**	**98.75**	94.44	96.42
LLaVA-Med [4]	91.04	86.21	87.59	86.83	81.32	77.80	74.81	76.02	95.83	93.16	93.16	93.16
SlideChat [3]	91.34	86.11	87.92	86.97	82.15	79.15	75.58	77.02	93.75	88.68	91.88	90.16
Ours	**94.57**	**92.66**	**91.34**	**91.85**	**85.79**	**85.03**	**79.45**	**81.70**	**97.92**	95.00	**98.72**	**96.72**

9. Implementation Details

In the first stage, following Slidechat [3], the projector and slide encoder are set to be trainable while the remaining components are frozen. This stage utilizes the WSI-caption data for initial alignment, employing a learning rate of 1e-3 for 3 epochs. In the second stage, we train the entire model for 2 epochs with a learning rate of 1e-4 and a batch size of 1. Specifically, we apply LoRA to the LLM to ensure efficient parameter updates. For the hyperparameters for our group sampler and patch selector, we employ a linear warmup schedule for the first 5000 iterations. During this warmup, the β_g weight increases from 0 to 0.1, and the β_p weight increases from 0 to 0.2, then they are held constant.

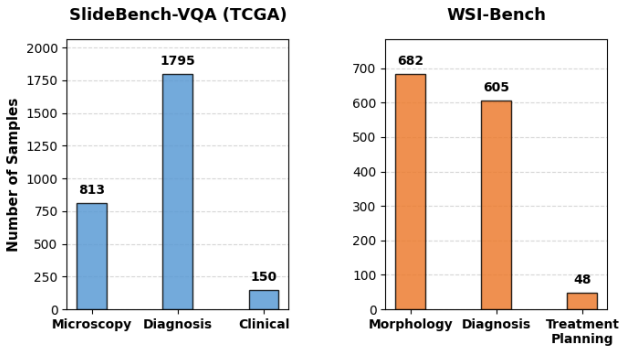

Figure 7. The sample distribution for the test set.

10. Additional Quantitative Results.

Figure 7 illustrates the sample distribution across different task categories in SlideBench-VQA and WSI-Bench. To further validate the reliability of HistoSelect, we report detailed macro-averaged metrics (Macro-Precision, Macro-Recall, and Macro-F1) across these two major benchmarks. As shown in Table 7 and Table 8, our method consistently achieves superior performance in these balanced metrics on both SlideBench-VQA and WSI-Bench. These improvements demonstrate that our approach effectively identifies task-relevant patches and generalizes well across various categories.

11. Additional Qualitative Results

To provide a more comprehensive evaluation of our proposed method, we present additional qualitative visualizations in this section. Due to the page constraints of the main manuscript, we extend our analysis here to demonstrate the model's performance across different data distributions. Specifically, we visualize the effectiveness of our question-aware selection mechanism on both the public TCGA dataset and an in-house Ovarian dataset. These results further validate that our method can consistently filter out background noise and diagnostically irrelevant patches, enabling the model to focus efficiently on the regions most related to the VQA query.

11.1. Results on Public Dataset

Figure 8 showcases the visualization results on the public TCGA dataset from WSI-Bench [5]. Consistent with the findings in the main text, our model demonstrates strong generalization capabilities. It successfully identifies and retains key histological features required to answer the question while discarding a significant portion of irrelevant and redundant patches, thereby ensuring that the downstream reasoning is primarily driven by the most informative patches.

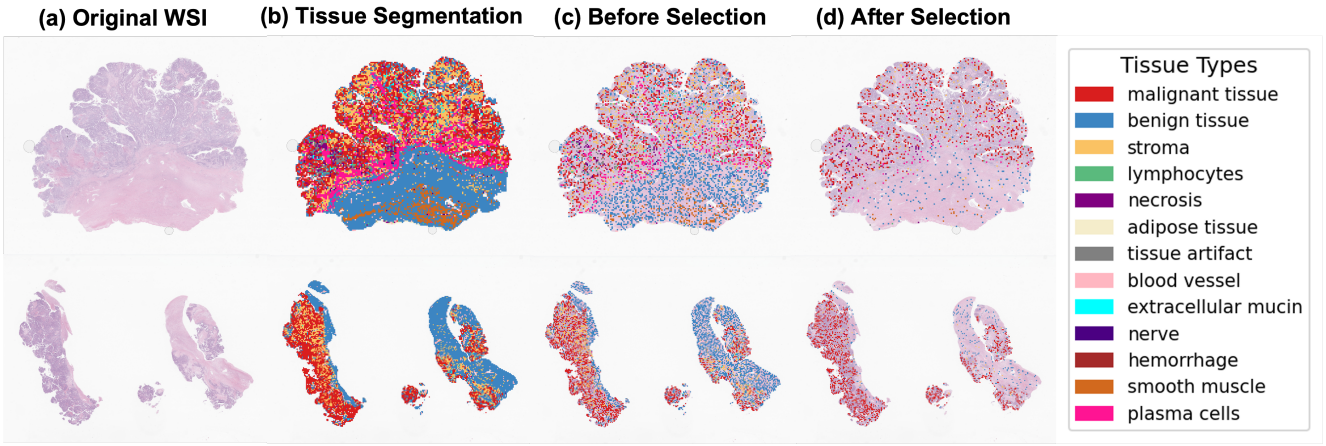

Figure 8. Additional visualization of the selection process on the public WSI-Bench dataset. (a) Original WSI. (b) The tissue segmentation mask. (c) A visualization of candidate patches extracted from tissue regions prior to selection. (d) The sparse set of patches retained by our model. As observed in (d), the model effectively suppresses irrelevant regions, focusing the attention solely on the informative patches required for the VQA task.

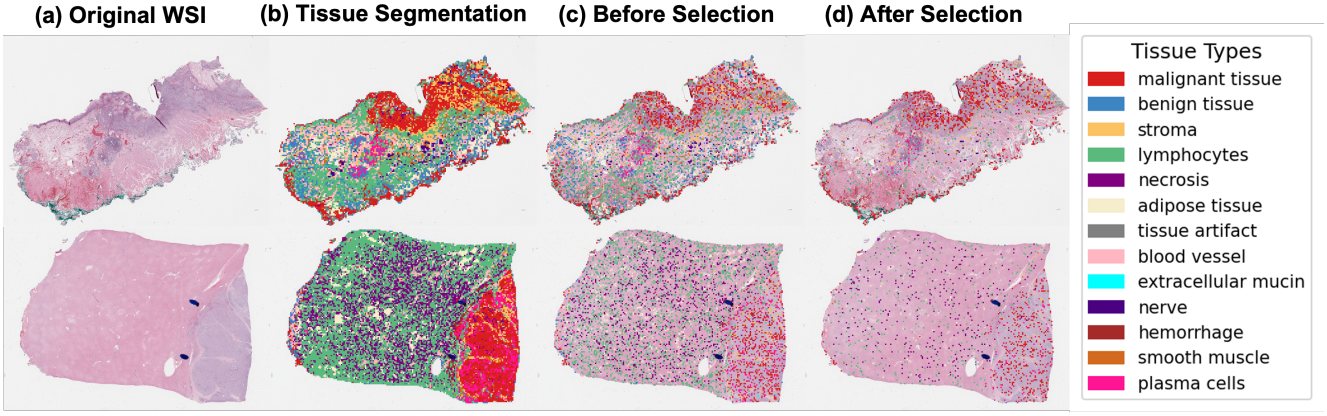

Figure 9. Visualization of the selection process on the private Ovarian dataset. The figure follows the same pipeline as the main manuscript and Figure 8: (a) Original WSI. (b) Tissue segmentation mask. (c) Patches before selection. (d) Patches after selection. These results demonstrate the robustness of our method against domain shifts common in private clinical data. The model successfully filters out non-informative tissue, preserving only the regions essential for accurate question answering.

11.2. Results on Private Dataset

To assess the robustness of our model in a real-world clinical setting, Figure 9 illustrates the selection process on our private ovarian dataset. Despite potential domain shifts such as variations in staining protocols and scanner properties compared to the public dataset, our question-aware selector maintains high precision. It effectively selects informative regions relevant to the query, verifying the method's applicability to proprietary clinical workflows.

11.3. Sampling Rate Distribution Analysis

To investigate the sampling rate distribution across different questions, we conducted a quantitative analysis on the Diagnosis and Morphology subset of the WSI-Bench dataset [5]. Specifically, we represented the tissue group sampling rate for each question as a 13-dimensional vector, where each dimension corresponds to the normalized sampling rate of a specific tissue component. We then applied K-means clustering to these vectors and visualized the resulting groupings using t-SNE, as shown in Figure 11. The emergence of four distinct clusters ($K = 4$) demonstrates that our model generates diverse and structured sampling patterns in the feature space. This grouping behavior confirms that the selection mechanism effectively navigates the complex composition of WSIs by adaptively prioritizing different histological tissue types based on the semantic focus of the question, rather than collapsing into a fixed, question-agnostic distribution.

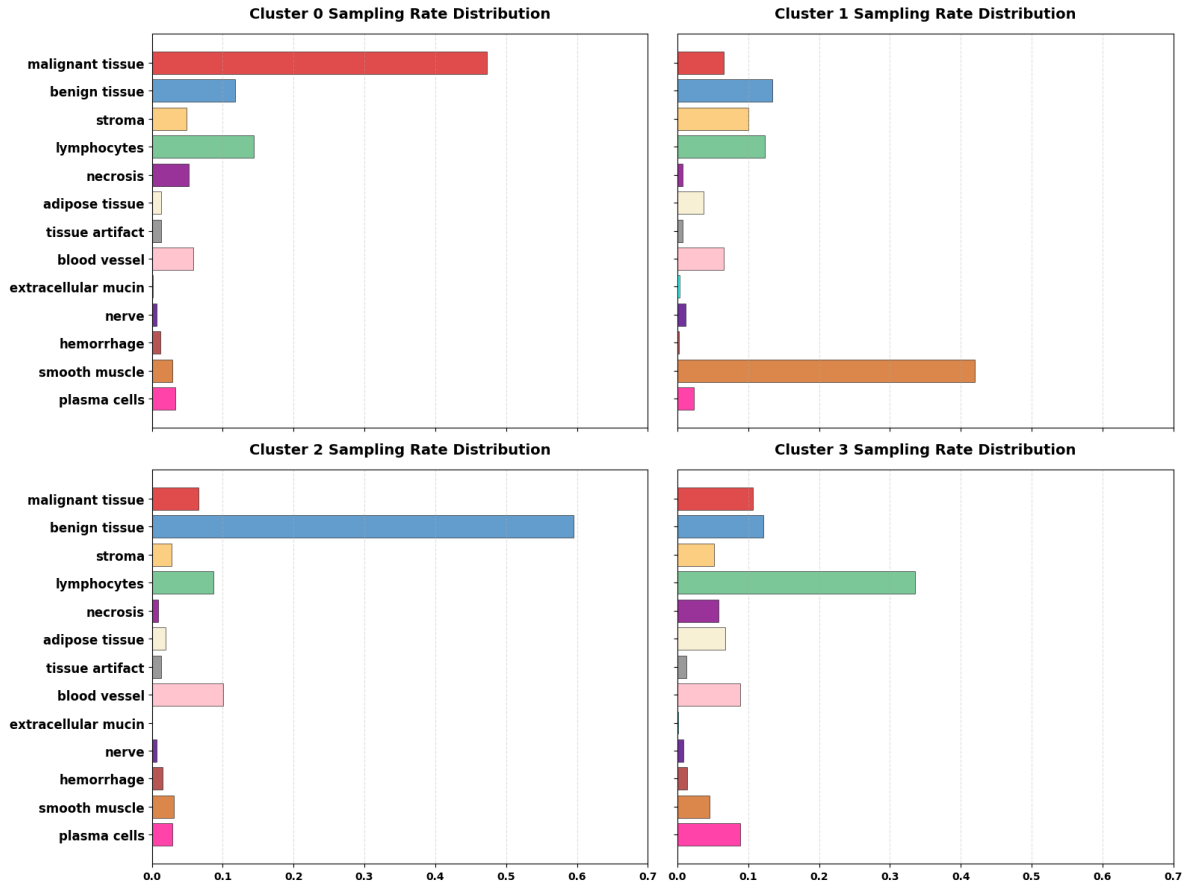

Figure 10. Mean sampling rate distribution bar charts for identified clusters. We report the average 13-dimensional sampling vectors for the four clusters discovered in Fig. 11. Each cluster exhibits a unique sampling rate pattern.

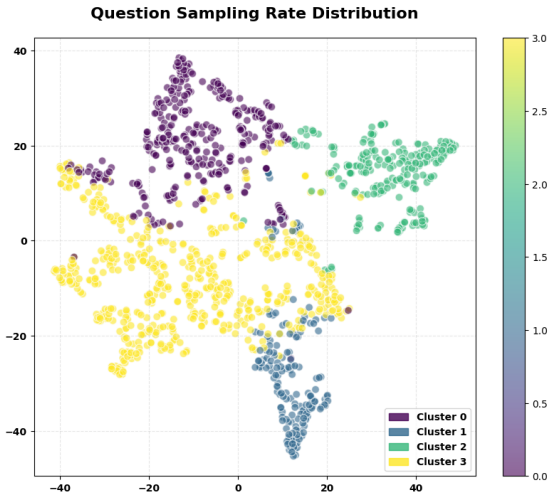

Figure 11. Visualization of question-aware sampling distributions. We visualize the 13-dimensional sampling rate vectors from the WSI-Bench Diagnosis and Morphology test set using t-SNE.

The cluster-specific sampling distributions, as visualized in Fig. 10, illustrate that our model develops distinct, task-driven sampling patterns. Each cluster corresponds to a specific clinical focus in the questions:

- **Cluster 0 (Tumor Classification)**: Prioritizes malignant tissue, aligning with questions focused on histological tumor grading and classification.
  Example: Based on the observed features, what do you think is the correct histological classification of the tumor? A) Adenocarcinoma B) Small cell carcinoma C) Squamous cell carcinoma D) Large cell carcinoma
- **Cluster 1 (Cellular Morphology)**: Shifts focus toward smooth muscle and stromal components, providing the necessary context for evaluating cellular variability and mitotic activity.
  Example: What are the notable features of the cellular morphology in this slide? A) Nuclei are uniform in appearance, showing no signs of active division. B) There is minimal variability in nuclear size, with a low rate of cell division. C) Nuclei appear extremely pleomorphic, with a very high rate of mitotic activity. D) There is moder-

ate variability in nuclear size and shape, with a moderate rate of cell division and presence of single cells.

- **Cluster 2 (Tissue Architecture)**: Shows a strong preference for benign tissue and surrounding structures, which is essential for assessing the overall microanatomy and glandular patterns.
  Example: What observations can you make about the tissue architecture on this slide? A) The tumor forms well-organized acinar structures with a clear glandular pattern. B) The tumor is characterized by prominent chicken-wire vasculature providing stroma. C) Tumor cells create extensive solid sheets, with a completely homogeneous pattern. D) The tissue maintains normal microanatomy with minimal deviation.
- **Cluster 3 (Tumor Infiltration)**: Concentrates on lymphocytes and extracellular components, capturing the critical interface where tumor cells infiltrate the stroma and adipose layers.
  Example: What is the observed pattern of tumor infiltration in this specimen? A) Tumor cells are limited to the submucosal layer without muscularis propria involvement. B) Tumor cells infiltrate the stroma, extending into the muscularis propria and adipose tissue. C) Tumor cells remain within glandular structures without stromal invasion. D) There is only infiltration into the adipose tissue, sparing the submucosal layer.

This clear divergence confirms that our selection mechanism is question-aware. Instead of relying on a static saliency map, the model dynamically re-prioritizes different histological tissue types based on the semantic intent of the question, ensuring that the most relevant patches are selected for each specific question.

12. Ablation Study

In this section, we conduct extensive ablation studies to evaluate the effectiveness of the proposed components in HistoSelect. We first analyze the impact of hyperparameters β_g and β_p in our loss function, which control the information bottleneck at the *group sampler* and *patch selector* levels, respectively. Subsequently, we investigate the influence of different training strategies and assess the model-agnostic generalization of our selector modules across various base models. Finally, a group selection analysis is performed to demonstrate the critical role of group-level selection in handling complex multi-tissue reasoning tasks.

β_g	Morphology	Diagnosis	Treatment
0	91.78	81.82	95.83
0.1	93.39	84.13	95.83
0.2	**94.57**	**85.79**	**97.92**
0.3	93.10	83.25	93.75

Table 9. Ablation Study on the weight β_g for the group sampler.

Impact of β_g for the Group Sampler. Table 9 reports the model performance under different values of $\beta_g \in \{0, 0.1, 0.2, 0.3\}$ while keeping β_p fixed. We observe that when $\beta_g = 0$, the group sampler lacks the necessary regularization to filter out irrelevant tissue groups, leading to a lower signal-to-noise ratio and suboptimal performance. As β_g increases to 0.2, the model effectively suppresses background noise at the group level, achieving the best overall accuracy across all tasks. However, further increasing β_g to 0.3 results in a performance drop. This suggests that an overly aggressive penalty causes the sampler to excessively reduce the sampling rate of tissue groups that contain necessary contextual information.

β_p	Morphology	Diagnosis	Treatment
0	92.07	81.32	93.75
0.05	93.25	84.23	95.83
0.10	**94.57**	**85.79**	**97.92**
0.15	93.83	83.74	95.83

Table 10. Ablation study on the weight β_p for the patch selector.

Impact of β_p for the Patch Selector. Table 10 examines the effect of the patch-level weight $\beta_p \in \{0, 0.05, 0.10, 0.15\}$. Similar to the group level, setting $\beta_p = 0$ tends to retain a large number of redundant patches, which introduces potential interference for the answer predictor. We find that setting $\beta_p = 0.10$ yields the optimal balance, allowing the model to identify the most distinct and discriminative patches without losing critical information. Conversely, setting β_p too high (e.g., 0.15) leads to over-pruning, where the model is penalized for retaining informative patches, causing a loss of fine-grained details essential for accurate diagnosis and treatment prediction.

Training Strategy	Morphology	Diagnosis	Treatment	Avg.
Joint Training	94.57	85.79	97.92	92.76
Joint + Patch Selector	94.71	85.95	97.92	92.86
Joint + Group Sampler	**95.15**	**86.11**	**97.92**	**93.06**

Table 11. Ablation on training strategies.

Impact of Training Strategies. Beyond hyperparameter tuning, we investigate whether extending the training procedure further impacts performance. As shown in Table 11, while the initial joint training yields strong results, conducting an additional epoch of training specifically for the sampler or selector modules proves beneficial. In particular, performing one extra epoch for the group sampler achieves the highest performance across most tasks. This suggests that after the joint training stage has established a solid foundation, the group sampler can further benefit from a dedicated optimization phase.

Ablation with Different Base Models. To verify the model-agnostic effectiveness of HistoSelect, we conduct an

experiment using Gemini 3 Flash as a frozen reasoning engine. Specifically, we randomly sample 200 cases from WSI-Bench and compare two input strategies: (1) a baseline using 100 randomly sampled patches with the question, and (2) using the top-100 patches selected by HistoSelect with the same question. As shown in Table 12, our method yields a consistent performance boost even for a stronger foundation model. This demonstrates our model's ability to filter redundant noise and identify task-relevant tokens independently of the base model's reasoning capacity.

Method	ACC	Macro-P	Macro-R	Macro-F1
Gemini 3 Flash	59.5	57.0	60.0	57.0
Gemini 3 Flash + HistoSelect	**62.5**	**58.0**	**64.0**	**59.0**

Table 12. Performance comparison on WSI-Bench (200 samples).

Impact of Group Selection. We further conduct an analysis by comparing our base model with a version using "Ideal Group Selection" (i.e., perfect identification of relevant tissue regions). As reported in Table 13, better group selection consistently leads to higher performance. Notably, the gain is more significant for multi-tissue questions (e.g., tumor infiltration patterns) compared to single-tissue ones (e.g., tumor detection). This highlights that the group sampler is particularly essential for handling complex clinical reasoning that requires cross-tissue contextual integration.

Question Type	Base	Base + Ideal Group	Gain (Δ)
Single-tissue (Easy)	85.0	87.0	+2.0
Multi-tissue (Hard)	73.0	79.0	+6.0

Table 13. Ablation on group selection.

13. Limitation

While our proposed method demonstrates promising results and improved efficiency in histopathology VQA, we acknowledge several limitations that outline directions for future research.

Evaluation on Other Datasets. First, our current experimental validation primarily focuses on the TCGA dataset and our in-house private dataset. While this covers a significant amount of variation, the heterogeneity of pathological data across different organs and scanning protocols is vast. To further verify the generalizability of our model, we intend to extend our training and testing to other large-scale public datasets, such as the BCNB (Early Breast Cancer Core-Needle Biopsy) [7] dataset. Evaluating on such diverse cohorts will help ensure our method remains robust across different cancer subtypes and data distributions.

Lack of Explicit Textual Reasoning. Second, while our method offers visual interpretability by highlighting the selected question-relevant patches, it does not currently generate explicit textual explanations justifying *why* these patches were selected. Providing a natural language rationale alongside the final VQA answer would further enhance trust in clinical decision-support systems. We aim to explore the integration of LLMs more deeply in future iterations to bridge this gap between visual attention and semantic reasoning.

References

[1] Streamlit. https://streamlit.io/, 2025. 2

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017. 1

[3] Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In *CVPR*, 2025. 4

[4] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS*, 2023. 4

[5] Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, et al. Wsi-llava: a multimodal large language model for whole slide image. In *ICCV*, 2025. 4, 5

[6] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *CVPR*, 2024. 4

[7] Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in oncology*, 2021. 8