Hepato-LLaVA: An Expert MLLM with Sparse Topo-Pack Attention for Hepatocellular Pathology Analysis on Whole Slide Images

Yuxuan Yang* [13], Zhonghao Yan* † [13], Yi Zhang* [2], Bo Yun[13], Muxi Diao[13], Guowei Zhao[2], Kongming Liang‡ [13], Wenbin Li‡ [2], Zhanyu Ma[13]

[1] Beijing University of Posts and Telecommunications
[2] National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College
[3] Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance
{wssf51379, zhonghao.yan}@bupt.edu.cn, zhangyi@cicams.ac.cn
*Equal Contribution †Project Lead ‡Corresponding Author

Abstract. Hepatocellular Carcinoma diagnosis relies heavily on the interpretation of gigapixel Whole Slide Images. However, current computational approaches are constrained by fixed-resolution processing mechanisms and inefficient feature aggregation, which inevitably lead to either severe information loss or high feature redundancy. To address these challenges, we propose Hepato-LLaVA, a specialized Multi-modal Large Language Model designed for fine-grained hepatocellular pathology analysis. We introduce a novel Sparse Topo-Pack Attention mechanism that explicitly models 2D tissue topology. This mechanism effectively aggregates local diagnostic evidence into semantic summary tokens while preserving global context. Furthermore, to overcome the lack of multi-scale data, we present HepatoPathoVQA, a clinically grounded dataset comprising 33K hierarchically structured question-answer pairs validated by expert pathologists. Our experiments demonstrate that Hepato-LLaVA achieves state-of-the-art performance on HCC diagnosis and captioning tasks, significantly outperforming existing methods. Our code and implementation details are available at https://github.com/PRIS-CV/Hepto-LLaVA.

Keywords: Hepatocellular Carcinoma · Multi-modal Large Language Models · Whole Slide Images · Digital Pathology

1 Introduction

Hepatocellular Carcinoma (HCC), a leading cause of global cancer mortality, relies on histopathological Whole Slide Images (WSIs) examination as the gold standard. However, HCC's high heterogeneity and complex microenvironment render gigapixel WSIs interpretation labor-intensive. Manual analysis is prone to inter-observer variability, especially for subtle early-stage lesions. Current Multiple Instance Learning (MIL) approaches [3] [10] [12] [18] are often limited

to classification, lacking clinical diagnostic capacity. This has catalyzed WSI-based Multi-modal Large Language Models (MLLMs) [9] [17] [2] [7] [15], which align visual features with language to enable Visual Question Answering (VQA).

A key design choice in these MLLMs is how to represent gigapixel WSIs for MLLMs. Existing methods typically fall into two categories: thumbnail-based approaches [9] [17] that resize the gigapixel image into a megapixel image, and slide-encoder-based approaches [2] [7] that aggregate thousands of patches into global tokens. However, thumbnail-based approaches inevitably lose patch-level details, while both methods ingest only WSIs inputs, consequently less equipped with patch-level capabilities essential for diagnosis. We investigate two primary research questions regarding the slide encoder, which acts as a bottleneck:

- **RQ1:** How can the slide encoder compress HCC WSIs into representations that preserve critical diagnostic details while minimizing redundancy?
- **RQ2:** How can the slide encoder accommodate variable-resolution inputs to generate features for multi-scale diagnosis in complex liver tissues?

To address **RQ1**, we design a **Sparse Topo-Pack Attention** mechanism that simulates the local aggregation and global juxtaposition of pathological diagnosis. By modeling 2D tissue topology, the slide encoder extracts topology-informed representations while significantly reducing spatial redundancy. A light-weight connector then condenses these features into fixed-length query tokens, effectively bridging the modality gap without losing critical diagnostic details. To address **RQ2**, we leverage hierarchical features to develop **HepatoPathoVQA**, a comprehensive multi-scale VQA dataset specifically for HCC. This dataset contains over 33K expert-validated QA pairs covering the full clinical workflow across three distinct scales: WSIs, Region of Interest (ROI, 5× mag.), and Patch (10× and 20× mag.). Finally, through Low-Rank Adaptation (LoRA) fine-tuning on this unified framework, we present **Hepato-LLaVA**, a specialized MLLM optimized for fine-grained hepatocellular pathology analysis.

In summary, our three main contributions are:

1. We construct **HepatoPathoVQA**, the first multi-scale WSI dataset for HCC, comprising over 33K QA pairs across three scales, which improves multi-scale modeling and bridges data with real-world clinical practice.
2. We introduce a **Sparse Topo-Pack Attention** mechanism that models 2D tissue topology, extracting global-context-aware representations for WSIs while mitigating information redundancy.
3. We present **Hepato-LLaVA**, a specialized MLLM optimized via a three-stage pipeline, achieving an improvement of 20% in average diagnostic accuracy in HCC over existing open-source pathology MLLMs.

2 Methods

2.1 Preliminary

Patch Encoding. Let $\mathbf{I}_{raw} \in \mathbb{R}^{H_{raw} \times W_{raw} \times 3}$ denote raw WSIs. We tessellate the tissue region into non-overlapping patches of size $P \times P$. This forms a grid

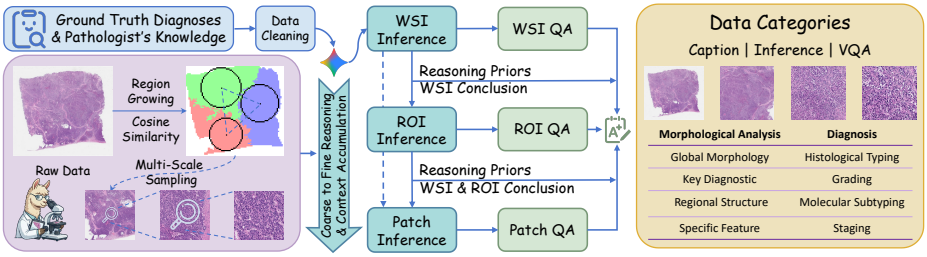

Fig. 1. Overview of the **HepatoPathoVQA** construction pipeline: (1) Extracts ROIs and Patches from WSIs using MST-based clustering and triangular seed-point selection. (2) Employs Gemini-3-flash for hierarchical inference by integrating macroscopic descriptions as context for subsequent microscopic analysis. (3) Generates multi-scale QA pairs and captions for instruction tuning and alignment.

layout of dimensions $H \times W$, where $H = \lfloor H_{raw}/P \rfloor$ and $W = \lfloor W_{raw}/P \rfloor$. Let $x_{i,j}$ represent the patch at the spatial coordinate (i, j) for $1 \leq i \leq H, 1 \leq j \leq W$. A frozen feature encoder f_{enc} then maps each patch into a D-dimensional embedding space, yielding a feature grid:

$$\mathbf{H}_{grid} = \left\{ \mathbf{h}_{i,j} \in \mathbb{R}^D \mid \mathbf{h}_{i,j} = f_{enc}(x_{i,j}) \right\}_{1 \leq i \leq H,\, 1 \leq j \leq W} \tag{1}$$

Slide Encoding. Since the raw feature $\mathbf{H}_{grid}$ contains thousands of patches, directly utilizing it for slide-level tasks introduces excessive dimensionality and redundant information. To address this, a slide encoder $\mathcal{T}(\cdot)$ is employed to extract semantic features from the flattened patch sequence $\mathbf{S}_{patch} = \text{Flatten}(\mathbf{H}_{grid})$:

$$\mathbf{H}_{slide} = \mathcal{T}(\mathbf{S}_{patch}) \tag{2}$$

In conventional MIL, $\mathcal{T}$ acts as a global aggregation function, where the output $\mathbf{H}_{slide} \in \mathbb{R}^{1 \times D}$ is a single compact token. In contrast, for Transformer-based approaches, $\mathcal{T}$ typically functions as a self-attention network where the output $\mathbf{H}_{slide} \in \mathbb{R}^{N \times D}$, often requiring a subsequent pooling to obtain a global token.

2.2 HepatoPathoVQA Dataset

Dataset Construction. We collected 200 WSIs containing HCC, with all patient identifiers removed to ensure privacy. From these WSIs, we constructed **HepatoPathoVQA**, a multi-scale dataset featuring 33K QA pairs centered on morphological analysis and diagnosis. As shown in Fig. 1, following the standard diagnostic workflow of pathologists, we developed a data generation pipeline using Gemini-3-flash [4]. This pipeline simulates the clinical reasoning process, transitioning from macroscopic observation to microscopic details. Each step in the pipeline is designed as an atomic operation. The data construction pipeline consists of three stages. **(1) Hierarchical Multi-Scale Sampling.** WSIs serve as the raw input. We identify ROIs using a triangular seed-point algorithm and

a Minimum Spanning Tree (MST). The MST aggregates adjacent patches based on cosine similarity to form three candidate regions. These ROIs represent a 5× magnification based on medical priors of cell adjacency. We manually remove uninformative regions. Finally, 10× and 20× local patches are randomly sampled around the ROI centers. **(2) Context-Injected Reasoning Generation.** Since Gemini cannot process gigapixel images, all regions are resized to 2048×2048 pixels. We design scale-specific prompts based on clinical focus and diagnostic status. Gemini generates descriptions in a hierarchical manner. Descriptions from higher scales serve as context for the next scale to ensure logical consistency from macroscopic to microscopic levels. **(3) Scale-aware Data Synthesis.** Leveraging the generated hierarchical descriptions, we generate diverse multi-scale QA pairs for instruction tuning. These pairs cover all three resolutions, enabling the model to perform both granular morphological analysis and holistic clinical diagnosis. Furthermore, these descriptions are utilized to produce high-quality image-text captions, which serve as the foundation for connector pre-training to achieve robust vision-language alignment.

Dataset Statistics. **HepatoPathoVQA** comprises 33,332 QA pairs. At our three proposed spatial scales, we follow the WSI-LLaVA [7] setup to prompt the model to generate reasoning chains along two axes—morphology and diagnosis. Each dimension is further subdivided into four subcategories, thereby supporting hierarchical diagnosis from overall patterns to cellular details. Beyond the VQA pairs, we provide **HepatoPathoCaption** (3,288 pairs) for alignment pre-training and **HepatoPathoBench** (3,056 pairs) split from HepatoPathoVQA for evaluation. The training and validation sets were strictly separated at the patient level. The entire dataset construction process was validated by three expert pathologists under a blind evaluation protocol, yielding high inter-rater consistency (Pearson $r = 0.96$) and a low rejection rate ($< 2\%$).

2.3 Sparse Topo-Pack Attention

Topological Mismatch in Existing Modeling. Existing pathology-specific MLLM predominantly rely on Longnet to extract features, treating the WSIs as a flattened 1D sequence. This assumption neglects the intrinsic 2D topological properties of pathological tissues. Biologically, semantic dependencies vary significantly across scales: local regions share a common semantic meaning(e.g., tumor margins crossing multiple patches), whereas distant regions are distinct with lower semantic coupling. To address this, we propose a hierarchical sparse attention mechanism that restores topological priors (Fig. 2, upper panel).

Hierarchical Sequence Construction. Unlike standard flattening, we construct a structured input sequence that explicitly encodes the grid hierarchy. We organize this grid into M Summary Packs in row-major order. Each pack $\mathbf{P}_m$ represents a local $k \times k$ window. We generate the pack summary token $\mathbf{s}_m$ by resizing the local image region of pack m and passing it through the patch encoder f_{enc} for semantic anchors. Similarly, the global token $\mathbf{g}_{global}$ is derived

by resizing entire WSIs and encoding it. The total sequence $\mathbf{S}_{in}$ is formed as:

$$\mathbf{S}_{in} = [\ \mathbf{g}_{global},\ \underbrace{\mathbf{h}^1_{1,1}, \ldots, \mathbf{h}^1_{k,k}, \mathbf{s}_1}_{\text{Pack } \mathbf{P}_1},\ \ldots,\ \underbrace{\mathbf{h}^M_{1,1}, \ldots, \mathbf{h}^M_{k,k}, \mathbf{s}_M}_{\text{Pack } \mathbf{P}_M}\ ] \tag{3}$$

where $\mathbf{h}^m_{i,j}$ represents the fine-grained patch tokens within the m-th pack. In our implementation, we set $k = 3$, resulting in 10 tokens per pack. Through the hierarchical attention mechanism, the summary token functions as a dynamic query that aggregates diagnostic evidence from patch tokens $\mathbf{h}^m_{i,j}$.

Hierarchical Sparse Mask. We define a hierarchical mask $\mathcal{M}$ to enforce the specific interaction rules. Let Ω_P and Ω_S denote the sets of all patch tokens and all summary tokens, respectively. Let $\mathcal{P}(t)$ denote the pack index of a token t. For any query token t_i and key token t_j in $\mathbf{S}_{in}$, the attention mask $\mathcal{M}_{i,j}$ is:

$$\mathcal{M}_{i,j} = \begin{cases} 0 & \text{if } t_j = \mathbf{g}_{global} \quad \text{(Global Sink)} \\ 0 & \text{if } t_i, t_j \in \Omega_P \wedge \mathcal{P}(t_i) = \mathcal{P}(t_j) \quad \text{(Intra-Pack Dense)} \\ 0 & \text{if } t_i \in \Omega_S \wedge t_j \in \Omega_P \wedge \mathcal{P}(t_i) = \mathcal{P}(t_j) \quad \text{(Aggregation)} \\ 0 & \text{if } t_i, t_j \in \Omega_S \quad \text{(Summary-Level Interaction)} \\ -\infty & \text{otherwise} \end{cases} \tag{4}$$

This hierarchical architecture effectively models pathological topology by harmonizing dual-scale interactions. The global token provides a macro-level reference, enabling local windows to focus on aggregating dense features into summary tokens. Summary tokens then facilitate long-range modeling, preserving the structural integrity of tissue regions across entire slide. In our implementation ($k = 3$), this sparsity reduces the theoretical attention overhead to $\approx \mathbf{1}\%$ of the dense counterpart. The end-to-end feature extraction takes 4.3 minutes per slide on average, compared to the 19.3 mins/slide required by Prov-GigaPath [13].

2.4 Hepato-LLaVA

MAE Pretrain. As shown in the lower panel of Fig. 2, we applied MAE pretraining on a composite dataset consisting of TCGA [11] and our internal data to optimize feature extraction. The dataset was augmented via spatial chunking and geometric flipping. The core innovation lies in a two-stage curriculum masking strategy: **Phase 1** implements patch-wise masking to reconstruct missing patches within intact packs, focusing on capturing tissue textures; **Phase 2** advances to pack-wise masking, where entire packs are removed to enforce long-range dependency modeling, thereby capturing high-level structural patterns.

MoCo Pretrain. Following MAE pre-training, this stage utilizes the Momentum Contrast (MoCo) [5] architecture. We utilize the un-augmented dataset to prevent the inclusion of augmented views as negative samples. This stage aims to align visual representations at the **Summary Token level**, focusing on tissue semantics rather than holistic descriptors. Addressing the high I/O costs of WSIs processing, we adapt the MoCo framework to operate on feature-level,

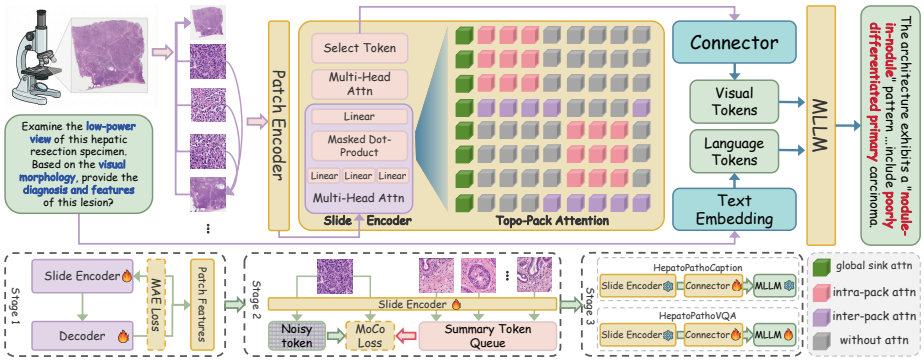

Fig. 2. Overview of the **Hepato-LLaVA** framework: (Upper) Incorporates Sparse Topo-Pack Attention into the model architecture. (Lower) Implements a three-stage training pipeline: MAE pre-training, MoCo pre-training, and instruction tuning (via LoRA). The sparse attention mask defines three topological interactions: (1) Global Sink for macro-context broadcasting, (2) Intra-Pack for local dense interactions, and (3) Inter-Pack for summary-level connections across packs.

generating positive pairs by injecting noise directly into token features. Coupled with a momentum-updated encoder and a queue containing cross-case negative samples, the model is optimized via InfoNCE loss to discriminate specific regions.

Multi-scale Instruction Tuning. To effectively bridge the modality gap between the pre-trained, frozen slide encoder and the MLLM, we introduce a Q-Former [6] Connector. To handle variable-length pack summary token sequences, this lightweight module employs a set of learnable query vectors to interact with inputs. The instruction tuning process is further divided into two sub-phases: (1) Visual-Language Alignment: We first train the connector on the **HepatoPath-Caption** dataset while keeping both the slide encoder and MLLM frozen. This stage aligns visual representations with linguistic semantics, enabling the connector to translate pathological features. (2) Diagnostic Instruction Tuning: We fine-tune both the connector and the MLLM on the **HepatoPath-VQA** dataset. This stage optimizes the model for diagnostic inference, enabling it to interpret multi-scale visual evidence and generate clinically precise responses.

3 Experiments

3.1 Experimental Settings

Datasets. The slide encoder is pre-trained on 10K WSIs aggregated from TCGA [11], HCMI [8], and internal dataset. For instruction tuning, we utilize HepatoPathoCaption (3K pairs) and HepatoPathoVQA (33K pairs). Evaluation is performed on HepatoPathoBench (3K pairs), strictly disjoint from the training.

Models. We conduct a comparison across three types of 7B models. For general medical models, we evaluate HuatuoGPT [16] and Lingshu [14]. For thumbnail-based pathology MLLMs, we include Quilt-LLaVA [9] and Patho-R1 [17](inputs

Table 1. Evaluation on HepatoPathoBench. Baselines include general medical MLLMs and pathology-specific MLLMs. Single/Multi refer to single/multiple choice tasks. WSI, ROI, and Patch denote the spatial scales. The Avg score is calculated across all tasks. Best results are **bold**; second-best are underlined

Model	Input	Morphological Analysis				Diagnosis				Multi-scale			Avg
		Open		Close		Open		Close					
		WSI-P	METEOR	Single	Multi	WSI-P	METEOR	Single	Multi	WSI	ROI	Patch	
Lingshu	Thumbnail	0.53	0.17	0.38	0.44	0.73	0.18	0.39	0.38	0.52	0.52	0.49	0.50
Huatuo-GPT	Thumbnail	0.74	0.24	0.81	0.45	0.70	0.23	0.59	0.32	0.60	0.65	0.65	0.65
Quilt-LLaVA	Thumbnail	0.64	0.22	0.47	0.32	0.56	0.15	0.57	0.37	0.57	0.60	0.55	0.57
Patho-R1	Thumbnail	0.66	0.19	0.87	0.50	0.20	0.05	0.59	0.45	0.55	0.55	0.54	0.55
SlideChat	WSI	0.70	0.17	0.87	0.47	0.72	0.14	0.63	0.39	0.66	0.68	0.66	0.66
WSI-LLaVA	WSI	0.69	0.20	0.84	0.46	0.67	0.16	0.65	0.36	0.65	0.67	0.64	0.65
Hepato-LLaVA	WSI	**0.79**	**0.33**	**0.97**	**0.88**	**0.75**	**0.33**	**0.87**	**0.68**	**0.82**	**0.83**	**0.83**	**0.83**

resized to their native square resolutions). For WSI-based pathology MLLMs with sparse attention, we compare against SlideChat [2] and WSI-LLaVA [7].

Metrics. To evaluate the clinical accuracy of open-ended responses, we employ METEOR [1] and WSI-P [7], which leverage LLM to evaluate predictions, providing a more reliable proxy for pathologist expertise. For single-choice questions, we adopt standard Accuracy based on a strict match. For multiple-choice questions, predictions that are correct but incomplete are assigned a score of 0.5.

3.2 Main Results

Hepato-LLaVA is initialized from WSI-LLaVA, which is fine-tuned using LoRA (r=128), while the connector is trained from scratch. We evaluated Hepato-LLaVA on HepatoPathoBench against six baselines. As shown in Table 1, while Thumbnail-based models yielded suboptimal results (Avg 0.50–0.57) due to severe information loss, Hepato-LLaVA achieved state-of-the-art performance (Avg **0.83**). This establishes a substantial lead of 0.17 over the best WSI-based runner-up SlideChat (0.66). In open-ended tasks, Hepato-LLaVA achieved a WSI-P of 0.79 in morphology and 0.75 in diagnosis, surpassing the SlideChat (0.70, 0.72). In close-ended tasks, our model attained 0.97 accuracy in morphological single-choice and 0.88 in multi-choice, significantly outperforming Patho-R1 (0.87, 0.50), a strong baseline enhanced by reinforcement learning. Regarding multi-scale consistency, Hepato-LLaVA demonstrated robust performance across WSI (0.82), ROI (0.83), and Patch (0.83) levels. This represents an improvement over our backbone model WSI-LLaVA (0.65, 0.67, 0.64), validating that the proposed sparse attention and connector effectively overcomes the scale variance limitations inherent in the base architecture. Qualitative evidence in Fig. 3 confirms its ability to interpret critical morphological evidence for precise diagnosis.

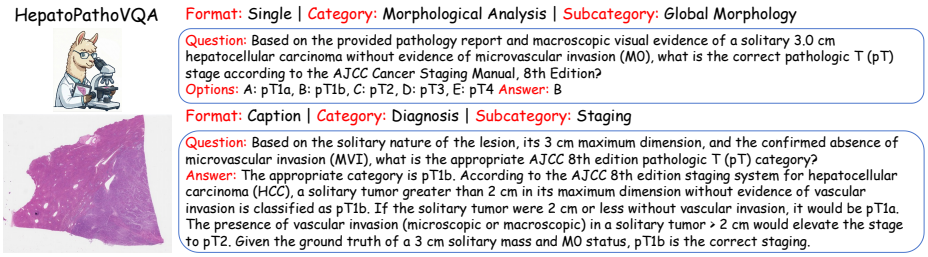

Fig. 3. A representative diagnostic case from HepatoPathoVQA, illustrating HepatoLLaVA's capability to provide precise by interpreting critical morphological evidence.

Table 2. Ablation of connector types and two-stage training strategy. For two-stage setting, we pre-train the connector on HepatoPathCaption with MLLM frozen.

Connector	Two-Stage	WSI	ROI	Patch
wsi-llava	✗	81.61	88.05	86.83
mlp	✗	74.43	82.83	79.97
mlp	✓	83.33	87.24	83.66
qformer	✗	**88.51**	87.82	85.34
qformer	✓	**88.51**	**88.75**	**88.10**

Table 3. Ablation of connector types and token numbers, reporting accuracy at different spatial scales (WSI/ROI/Patch).

Connector	Token Num	WSI	ROI	Patch
wsi-llava	all	81.61	88.05	86.83
mlp	all	83.33	87.24	83.66
mlp	32	85.34	**89.10**	86.13
mlp	1	85.34	87.47	87.25
qformer	500	**89.37**	88.40	85.25
qformer	32	88.51	88.75	**88.10**
qformer	1	87.64	88.52	86.65

3.3 Ablation Study

To validate the effectiveness of our architectural design, we conducted ablation studies on the HepatoPathoBench with identical hyperparameters.

Effectiveness of Connector and Training Strategy. We compared our Q-Former-based connector against the MLP used in baseline methods like WSI-LLaVA [7]. As shown in Table 2, the Q-Former architecture significantly outperforms the MLP baseline across all scales (+5.18% on WSI). It also exhibits superior stability ($std = 0.26$) compared to the MLP ($std = 1.77$), validating its effectiveness in resampling variable-length visual sequences into fixed query embeddings. Furthermore, the results indicate that the Two-Stage setting yields performance gains (+2.76% on ROI) compared to direct fine-tuning. We also report the performance of the MLP connector initialized from WSI-LLaVA weights, which benefits from broader pre-training and serves as a strong baseline.

Analysis of Feature Redundancy and Token Efficiency. We hypothesize (RQ2) that standard slide encoders generate excessively redundant representations from gigapixel WSIs. To validate this, we evaluated the impact of token quantity on model performance. As detailed in Table 3, increasing the token count introduces noise and degrades performance. The Q-Former with 32 learnable queries achieves peak accuracy (88.75% on ROI, 88.10% on Patch), outperforming the 500-token configuration. A similar trend is observed in the MLP baseline, where even single-token(via max pooling) surpasses the use of all pack tokens. These results corroborate the sparsity of diagnostic signals in WSIs

and underscore the necessity of condensing visual features into pack summaries. While the 500-query Q-Former excels in handling the extended sequences typical of WSIs, we select the 32-query setting to strike an optimal balance between inference efficiency and robust performance across varying spatial scales.

4 Conclusion

We present Hepato-LLaVA, a specialized MLLM for hepatocellular pathology analysis. To mitigate information redundancy in WSIs, we propose a Sparse Topo-Pack Attention mechanism that explicitly models 2D tissue topology. We also establish HepatoPathoVQA, a comprehensive multi-scale VQA dataset specifically for HCC WSIs. Extensive experiments demonstrate that our model achieves state-of-the-art performance, validating that topology-aware representations outperform high-dimensional redundant features. We believe this work demonstrates the efficacy of embedding pathological priors into deep learning frameworks, advancing the frontier of efficient AI for precision pathology.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant Nos. U23B2052, 62476029, 62225601, and 82541023), the National Key Research and Development Program of China (Grant No. 2026ZD0557402), the Beijing Natural Science Foundation (Grant No. JQ24045), the Fundamental Research Funds for the Beijing University of Posts and Telecommunications (Grant No. 2025TSQY08), and the Beijing Nova Program and the Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
2. Chen, Y., Wang, G., Ji, Y., Li, Y., Ye, J., Li, T., Hu, M., Yu, R., Qiao, Y., He, J.: Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5134–5143 (2025)
3. Fourkioti, O., De Vries, M., Jin, C., Alexander, D.C., Bakal, C.: Camil: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images. arXiv preprint arXiv:2305.05314 (2023)
4. Google: Gemini 3 flash (free tier). `https://gemini.google.com/` (2026), accessed: 2026-02-02

5. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9729–9738 (2020)
6. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
7. Liang, Y., Lyu, X., Chen, W., Ding, M., Zhang, J., He, X., Wu, S., Xing, X., Yang, S., Wang, X., et al.: Wsi-llava: A multimodal large language model for whole slide image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22718–22727 (2025)
8. National Cancer Institute, National Human Genome Research Institute, European Bioinformatics Institute: Human cancer models initiative (hcmi). `www.cancer.gov/ccg/research/functional-genomics/hcmi` (2017), accessed: 2026-02-03
9. Seyfioglu, M.S., Ikezogwo, W.O., Ghezloo, F., Krishna, R., Shapiro, L.: Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13183–13192 (2024)
10. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11566–11578 (2024)
11. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M.: The cancer genome atlas pan-cancer analysis project. Nature genetics **45**(10), 1113–1120 (2013)
12. Wu, J., Chen, M., Ke, X., Xun, T., Jiang, X., Zhou, H., Shao, L., Kong, Y.: Learning heterogeneous tissues with mixture of experts for gigapixel whole slide images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5144–5153 (2025)
13. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (2024)
14. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
15. Yan, Z., Diao, M., Yang, Y., Jing, R., Xu, J., Zhang, K., Yang, L., Liu, Y., Liang, K., Ma, Z.: Medreasoner: Reinforcement learning drives reasoning grounding from clinical thought to pixel-level precision. arXiv preprint arXiv:2508.08177 (2025)
16. Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Chen, G., Li, J., Wu, X., Zhiyi, Z., Xiao, Q., et al.: Huatuogpt, towards taming language model to be a doctor. In: Findings of the association for computational linguistics: EMNLP 2023. pp. 10859–10885 (2023)
17. Zhang, W., Zhang, P., Guo, J., Cheng, T., Chen, J., Zhang, S., Zhang, Z., Yi, Y., Bu, H.: Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner. arXiv preprint arXiv:2505.11404 (2025)
18. Zhuang, Z., Zhou, F., Wang, L.: Libra-mil: Multimodal prototypes stereoscopic infused with task-specific language priors for few-shot whole slide image classification. arXiv preprint arXiv:2511.07941 (2025)