



MLLM-HWSI: A Multimodal Large Language Model for Hierarchical Whole Slide Image Understanding

Basit Alawode¹, Arif Mahmood², Muaz Khalifa Al-Radi¹, Shahad Albastaki¹, Asim Khan¹, Muhammad Bilal³, Moshira Ali Abdalla¹, Mohammed Bennamoun⁴, Sajid Javed¹

¹Department of Computer Science, Khalifa University of Science and Technology, UAE.

²Information Technology University, Pakistan. ³KAU, KSA. ⁴University of the Western Australia.

Contents

1. Inference Details	1
2. Hierarchical WSI-Caption Alignment	1
3. Additional Ablation Studies	4
3.1. Cell Segmentation Backbones (Table 1)	4
3.2. Impact of Different Visual Encoders (Table 2)	4
3.3. Variants of ViT _{cell-cell} (Table 3)	4
3.4. Importance of Hierarchical Representations (Table 4)	4
3.5. Effect of the LLM (Table 5)	4
3.6. Semantic Patch Filtering (SPF) (Table 6-7)	4
4. Computational Complexity	5
5. WSI-level Classification Results	5
5.1. Zero-shot Classification of WSIs (Table 8)	5
5.2. Linear Probe Evaluation (Table 9)	5
6. WSI-Level Report Generation Qualitative Results (Tables 10-14)	6
7. Pre-training Details of MLLM-HWSI	6
8. Computational Pathology Datasets	7

1. Inference Details

Each WSI is partitioned into ≈ 20 regions, each with 256 patches. Since SPF has two components: (i) HPS (Eq. 1), which removes redundant patches using visual similarity only, and (ii) DPS (Eq. 2), which leverages report-derived semantic tokens to guide patch relevance during training. Therefore, during inference, no pathology reports are used. Only HPS is applied, so patch selection is fully vision-based with no test-time information leakage.

2. Hierarchical WSI-Caption Alignment

In Computational Pathology (CPath), the importance of hierarchical alignment arises from both biological reasoning and representational learning principles [17, 22, 43]. Theoretically, WSIs are not uniform visual entities; instead, they exhibit a nested organization, where meaning emerges across multiple levels of abstraction [11, 16, 26, 48]. Diagnostic semantics are inherently hierarchical: cellular morphology defines nuclear atypia and mitotic figures; patch-level structures capture gland formation, necrosis, or immune infiltration; region-level context reflects tumor invasion and stromal interaction; and the global WSI conveys architectural disarray and overall differentiation [4, 19, 32]. A single global embedding, as used in conventional MLLMs [13, 35], collapses this structure and causes information loss, particularly of the spatial and semantic dependencies that exist between local and global tissue organization. Hierarchical alignment mitigates this by learning distinct yet interconnected visual–language mappings for each scale. Each level aligns with its corresponding linguistic abstraction—cells correspond to morphological words, patches to descriptive phrases, regions to structural sentences, and the WSI to a diagnostic paragraph—thus preserving compositional semantics and ensuring that information propagates coherently across scales [6, 21, 30].

Therefore, the hierarchical WSI–caption alignment mechanism in MLLM-HWSI is central to connecting the visual semantics of histopathology with the descriptive reasoning expressed in diagnostic language [14, 18, 44]. In conventional CPath MLLMs [13, 35], caption alignment is performed only at the global level—linking an entire WSI to its corresponding report or summary. While effective for coarse labeling or WSI-level classification, this approach overlooks the fine-grained relationships between local morphological features and the textual phrases that describe them. Hierarchical WSI–caption alignment overcomes this limitation by establishing multi-level correspon-

Cell Segmentation Backbone	Cell Feat.	PANDA [7] (BA)	EBRAINS [42] (BA)	WSI-VQA [10] (A)	SlideBench-VQA (BCNB) [13](A)
CellViT [29]	ViT _{cell-cell}	0.748	0.612	0.692	0.687
NuHTC [33]	ViT _{cell-cell}	<u>0.733</u>	<u>0.600</u>	0.689	<u>0.685</u>
STRARDIST [45]	ViT _{cell-cell}	0.701	0.585	0.665	0.670
MicroNet [41]	ViT _{cell-cell}	0.698	0.561	0.672	0.671
HoverNet [23]	ViT _{cell-cell}	0.725	0.591	<u>0.690</u>	0.673

Table 1. **Effect of cell segmentation backbones in ViT_{cell-cell}.** Results show Balanced Accuracy (BA) for PANDA and EBRAINS, and Accuracy (A) for WSI-VQA and SlideBench-VQA (BCNB). CellViT achieves the highest scores, confirming the benefit of SAM-based segmentation for cell-level feature extraction.

Models	Cell Encoder	Patch Encoder	Region Encoder	WSI Encoder	PANDA [7] (BA)	EBRAINS [42] (BA)	WSI-VQA [10] (A)	SlideBench-VQA (BCNB) [13] (A)
MLLM-HWSI	ViT _{cell-cell}	$\mathcal{F}_{\text{CONCH}}$	ViT _r	ViT _{WSI}	0.748	0.612	0.692	0.687
MLLM-HWSI	ViT _{cell-cell}	UNI	UNI	UNI	0.721	<u>0.589</u>	0.653	0.664
MLLM-HWSI	ViT _{cell-cell}	CONCH	CONCH	CONCH	<u>0.712</u>	0.581	0.644	0.657
MLLM-HWSI	ViT _{cell-cell}	CONCH	CONCH	LongNet	0.702	0.575	<u>0.681</u>	0.665
MLLM-HWSI	ViT _{cell-cell}	GigaPath	GigaPath	LongNet	0.692	0.562	0.673	<u>0.681</u>
MLLM-HWSI	ViT _{cell-cell}	UNI	UNI	LongNet	0.686	0.564	0.663	0.634

Table 2. **Influence of visual encoder selection across hierarchical levels.** Different combinations of patch-, region-, and WSI-level encoders (UNI, CONCH, GigaPath, LongNet) are evaluated, all fine-tuned with the proposed loss. The $\mathcal{F}_{\text{CONCH}}$, ViT_r, and ViT_{WSI} configuration yields the best overall results, highlighting the importance of heterogeneous multi-scale encoders.

dences between visual evidence and linguistic descriptions across the full diagnostic hierarchy, enabling precise, interpretable, and clinically coherent visual–language reasoning.

At the representational level, hierarchical caption alignment ensures that visual embeddings from different hierarchical levels—cellular, patch-level, regional, and global—are aligned with language tokens of equivalent semantic granularity. Words or short phrases describing morphology (e.g., “hyperchromatic nuclei,” “mitotic figures”) align naturally with cell-level embeddings; sentences describing structural patterns (e.g., “disorganized glandular arrangement”, “stromal invasion”) align with region-level features; and full diagnostic summaries align with the WSI-level representation [11, 43]. This multi-scale correspondence transforms caption generation from a monolithic text synthesis problem into a structured reasoning process, where the model progressively integrates information across scales to compose a coherent narrative of pathology. The result is a caption that not only summarizes findings but also reflects how human pathologists articulate diagnostic observations.

From a clinical perspective, hierarchical WSI–caption alignment bridges the gap between machine perception and human explanation [27, 40]. In real-world diagnostic practice, pathologists document their findings hierarchically: starting with cellular morphology, describing archi-

tectural context, and concluding with a diagnostic impression [2, 6, 21, 28]. For example, a typical breast carcinoma report might read, “*The tumor displays irregular ductal structures lined by pleomorphic epithelial cells with hyperchromatic nuclei and increased mitotic activity.*” Each component of this description corresponds to a specific spatial scale within the tissue. By aligning these text segments with the respective visual features, MLLM-HWSI enables the model to “*speak the language of pathology*” — generating captions that explicitly refer to verifiable visual evidence. This interpretability enhances clinical transparency, allowing practitioners to trace each diagnostic statement back to its morphological basis, a critical requirement for medical AI adoption.

On a modeling level, hierarchical caption alignment serves as an additional supervisory signal that strengthens the multi-scale visual–language embedding space. Aligning visual tokens with hierarchical captions encourages the network to encode features that are both discriminative for diagnosis and descriptive for reporting. This dual objective reduces overfitting to classification labels and promotes a richer representation capable of supporting diverse downstream tasks, including report generation, retrieval, and VQA. Furthermore, the caption alignment process improves semantic calibration between local and global features: by ensuring that lower-level embeddings contribute meaning-

Variants	ViT _{cell-cell}			PANDA	EBRAINS	WSI-VQA	SlideBench-VQA
	# Encoder (n)	# heads (h)	Dimension (d)	(BA)	(BA)	(A)	(BCNB) (A)
a. MLLM-HWSI	2	2	768	0.748	0.612	0.692	0.687
b. MLLM-HWSI	4	4	768	0.726	0.592	0.681	<u>0.677</u>
c. MLLM-HWSI	6	6	768	0.727	0.590	0.682	0.675
d. MLLM-HWSI	2	2	384	<u>0.741</u>	0.595	0.688	0.671
e. MLLM-HWSI	2	2	192	0.723	<u>0.596</u>	<u>0.690</u>	0.676
Variants	Pooling Operation			PANDA	EBRAINS	WSI-VQA	SlideBench-VQA
	Max	Min	Average	(BA)	(BA)	(A)	(BCNB) (A)
f. MLLM-HWSI	✓			0.615	0.521	0.653	0.621
g. MLLM-HWSI		✓		0.605	0.545	0.636	0.618
h. MLLM-HWSI			✓	0.593	0.543	0.648	0.635

Table 3. **Effect of ViT_{cell-cell} architecture on performance.** Variants (a–e) modify the number of encoders (n), heads (h), and embedding dimensions (d), while (f–h) use max, min, and average pooling instead of attention. Results (BA for PANDA/EBRAINS, A for WSI-VQA/SlideBench-VQA) show that the $n = 2, h = 2, d = 768$ configuration performs best, emphasizing the value of attention-based cell-level modeling.

Models	Cell Feat.	Patch Feat.	Region Feat.	WSI Feat.	PANDA [7] (BA)	EBRAINS [42] (BA)	WSI-VQA [10] (A)	SlideBench-VQA (BCNB) [13] (A)
WSI-LLaVA [35]	×	×	×	✓	0.644	0.501	0.546	0.553
SlideChat [13]	×	×	×	✓	0.633	0.466	0.601	0.541
MLLM-HWSI ₁	×	×	×	✓	0.661	0.519	0.616	0.576
MLLM-HWSI ₂	×	×	✓	✓	0.686	0.534	0.611	0.592
MLLM-HWSI ₃	×	✓	✓	✓	0.711	0.566	0.661	0.621
MLLM-HWSI ₄	✓	×	×	✓	0.674	0.531	0.613	0.588
MLLM-HWSI ₅	×	✓	×	✓	0.698	0.548	0.623	0.606
MLLM-HWSI ₆	✓	✓	×	✓	<u>0.715</u>	0.575	<u>0.669</u>	0.640
MLLM-HWSI ₇	✓	×	✓	✓	0.714	0.587	0.668	<u>0.653</u>
MLLM-HWSI	✓	✓	✓	✓	0.748	0.612	0.692	0.687
MLLM-HWSI ₈	✓	×	×	×	0.616	0.476	0.569	0.522
MLLM-HWSI ₉	×	✓	×	×	0.623	0.491	0.578	0.521
MLLM-HWSI ₁₀	×	×	✓	×	0.631	0.511	0.581	0.529
MLLM-HWSI ₁₁	✓	✓	×	×	0.675	0.543	0.621	0.577
MLLM-HWSI ₁₂	✓	×	✓	×	0.672	0.538	0.618	0.574
MLLM-HWSI ₁₃	×	✓	✓	×	0.673	0.535	0.612	0.566
MLLM-HWSI ₁₄	✓	✓	✓	×	0.712	<u>0.588</u>	0.666	0.623

Table 4. **Effect of hierarchical representations in MLLM-HWSI.** Progressive inclusion of cell-, patch-, region-, and WSI-level features in MLLM-HWSI_{1–3} improves performance across all benchmarks. The full MLLM-HWSI achieves the highest scores, confirming the importance of hierarchical multi-scale alignment. PANDA and EBRAINS datasets are used for zero-shot classification while WSI-VQA and SlideBench-VQA (BCNB) datasets are used for VQA task. Feat. stands for “Features”, BA stands for “Balanced Accuracy”, and A stands for “Accuracy”.

fully to higher-level textual synthesis, the model maintains consistency between fine-grained details and WSI-level conclusions.

Empirically, hierarchical WSI-caption alignment enables MLLM-HWSI to produce captions that resemble expert-pathology reports—concise yet semantically dense, containing morphological detail, architectural context, and diagnostic interpretation in a single, coherent paragraph.

Such outputs demonstrate not only the model’s ability to describe what is visible but also to explain why those features are diagnostically relevant. This capability moves beyond simple visual description toward clinically useful, interpretable reasoning, establishing MLLM-HWSI as a bridge between computational pathology and real-world diagnostic reporting.

3. Additional Ablation Studies

3.1. Cell Segmentation Backbones (Table 1)

Table 1 reports the performance when the backbone cell segmentation method is varied within ViT_{cell-cell}. The SAM-based CellViT [29] achieves the best results.

3.2. Impact of Different Visual Encoders (Table 2)

Table 2 replaces patch/region encoders with UNI, CONCH, or GigaPath, and the WSI encoder with UNI, CONCH, or LongNet, using aggregation layers trained under our losses. Homogeneous stacks (all-UNI or all-CONCH) reduce feature diversity and underperform the proposed encoder mix. Combining LongNet with CONCH, GigaPath, or UNI improves over homogeneous variants but still lags our proposed configuration.

3.3. Variants of ViT_{cell-cell} (Table 3)

Table 3 studies architectural choices for ViT_{cell-cell}: number of encoder blocks $n \in \{2, 4, 6\}$, heads $h \in \{2, 4, 6\}$, and embedding dimension $d \in \{768, 384, 192\}$. The configuration $n=2, h=2, d=768$ yields the best overall results. Replacing ViT_{cell-cell} with simple min/max/average pooling leads to significant degradation, indicating the necessity of attention-based cell-cell interaction.

3.4. Importance of Hierarchical Representations (Table 4)

As shown in Table 4, we progressively augment the hierarchical features in MLLM-HWSI₁₋₃. Using only WSI-level features (MLLM-HWSI₁) already exceeds baseline methods. Adding region, patch, and cell-level features yields consistent improvements across all datasets. A complementary *subtractive* study (MLLM-HWSI₄₋₇) causes notable drops, underscoring the importance of every representation level.

Table 4 analyzes the contribution of hierarchical representations at different hierarchical levels within MLLM-HWSI. The variants MLLM-HWSI₁₋₃ incrementally incorporate additional levels of hierarchy—starting from WSI-level features alone, then progressively adding region-, patch-, and cell-level embeddings. Even with only WSI-level features (MLLM-HWSI₁), the model already surpasses strong baselines such as SlideChat and WSI-LLaVA, indicating that the hierarchical pre-training strategy captures rich global contextual features. As finer-scale information is introduced, performance consistently improves across all datasets. The proposed MLLM-HWSI model, which combines cell-, patch-, region-, and WSI-level embeddings, achieves the best overall performance, reaching 74.80% and 61.20% balanced accuracy on PANDA and EBRAINS, respectively, and 69.20% and 68.70% accuracy on WSI-VQA and SlideBench-VQA.

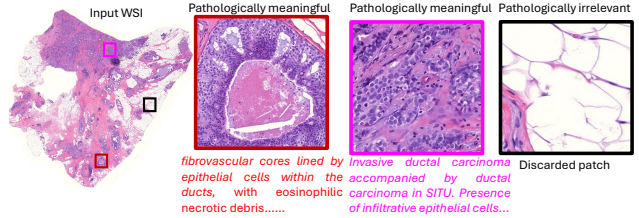


Figure 1. Pathologically meaningful patches and discarded patch.

These gains demonstrate that hierarchical representations allow the model to integrate cellular morphology, microarchitectural context, and global tissue organization into a unified reasoning process. The complementary subtractive analysis (MLLM-HWSI₄₋₇) further validates this effect—removing any representation hierarchy leads to a measurable drop in performance, particularly when cell- or patch-level features are excluded, reflecting the importance of fine-grained morphological grounding. Models retaining only cell-, patch-, or region-level features (MLLM-HWSI₈₋₁₀) perform significantly worse, underscoring the necessity of multi-scale contextual integration.

Overall, these results confirm that each hierarchical representation contributes meaningfully to diagnostic accuracy. The full MLLM-HWSI, which aligns all four levels of representation, yields the most robust and interpretable performance, emulating how pathologists synthesize information across magnifications—from cellular detail to WSI-level context—to reach precise diagnostic conclusions.

3.5. Effect of the LLM (Table 5)

Table 5 evaluates Vicuna-7B-v1.5 [15], Phi-3-Mini-4k-Instruct [1], Llama3-8B-Instruct [24], InternLM2-Chat-7B [8], and Qwen2-2.5 7B-Instruct [47] within MLLM-HWSI. Qwen2-2.5 7B-Instruct attains the best performance; the other four are competitive, highlighting the generalization of our framework.

3.6. Semantic Patch Filtering (SPF) (Table 6-7)

Table 6 examines Heterogeneous Patch Selection (HPS) and Diagnostically Relevant Patch Selection (DPS). For DPS we select top- $k=48$ patches per region R_i (Table 7). Removing HPS and/or DPS substantially degrades performance; substituting HPS with k -means clustering also reduces accuracy. Table 7 varies the DPS top- $k \in \{32, 64, 96\}$ (and additional values), with the best results at $k=48$. Pathologically meaningful qualitative patches are shown in Fig. 1.

In our experiments, SPF dynamically selects 48 patches per region (Table 6-7) before LLM input. After SPF and cell-cell attention fusion (ViT_{cell-cell}), each patch yields one cell and one patch token, each region yields one region token, plus one WSI token, resulting in ~ 1941 tokens/WSI

LLMs	SlideBench (BCNB) (A)	WSI- VQA (A)	PANDA (BA)
Vicuna-7B-v1.5 [15]	<u>0.682</u>	0.683	0.725
Phi-3-Mini-4k-Instruct [1]	0.664	0.674	0.706
Llama3-8B-Instruct [24]	0.677	0.673	0.725
Internlm2-Chat-7B [8]	0.681	<u>0.684</u>	<u>0.738</u>
Qwen2-2.5 7B-Instruct [47]	0.687	0.692	0.748

Table 5. **Effect of LLM choice on VQA performance.** Comparison of five instruction-tuned LLMs integrated into MLLM-HWSI across SlideBench (BCNB), WSI-VQA, and PANDA datasets. Qwen2.5-7B-Instruct yields the highest accuracy, highlighting its stronger multimodal reasoning capability.

always below the 2048 token limit, with no truncation. For a 4096-dim FP16 LLM, this corresponds to ~ 15 MB of input embeddings and ~ 30 – 45 MB total memory, including the KV cache.

4. Computational Complexity

The model was implemented on four NVIDIA A100 GPUs. During zero-shot inference, MLLM-HWSI required an average of 4.90 minutes per WSI on the BRAINS30 dataset, compared to 4.3, 4.4, and 3.8 minutes for SlideChat, TITAN, and WSI-LLaVA, respectively. The additional time arises from multi-scale feature extraction and semantic patch filtering, which enhance performance at a modest computational cost. *Despite incorporating hierarchical multi-scale feature extraction, MLLM-HWSI maintains computational efficiency comparable to existing SOTA models, demonstrating scalability without significant inference overhead.*

5. WSI-level Classification Results

5.1. Zero-shot Classification of WSIs (Table 8)

We evaluated the zero-shot WSI classification capability of the pre-trained MLLM-HWSI model using the vision and text encoders obtained from Stage I (hierarchical cross-modal alignment). Following established evaluation protocols in TITAN [20], CONCH [37], and QuiltNet [31], we directly measured the semantic alignment between hierarchical WSI features and class-specific textual descriptions without any task-specific fine-tuning.

For each test WSI, hierarchical visual features were extracted from the MLLM-HWSI encoder and compared against class-level textual prompts encoded by the text encoder. Both visual and textual embeddings were ℓ_2 -normalized, and class prediction was determined by selecting the label corresponding to the highest cosine similarity between the two modalities. We adopted dataset-specific testing prompts consistent with prior zero-shot WSI classification works to ensure fair comparison across benchmarks [12, 20, 31, 37].

This protocol evaluates how effectively MLLM-HWSI transfers its learned hierarchical alignment from multi-modal pre-training to unseen classification tasks. As shown in Fig. 4(a) of the main paper and Table 8, MLLM-HWSI achieves SOTA zero-shot accuracy across six external datasets, demonstrating robust generalization and the discriminative strength of its multi-scale visual–language representations.

5.2. Linear Probe Evaluation (Table 9)

We also conducted a linear probe evaluation to assess the discriminative strength and transferability of the representations learned by MLLM-HWSI during pre-training. Linear probing provides a widely adopted, architecture-agnostic framework for measuring the quality of learned features [12, 20]. The procedure involves freezing all parameters of the pre-trained encoder and training a simple logistic regression classifier on the extracted features. High linear probe performance indicates that the encoder captures rich, separable, and generalizable representations. Please see our linear probe evaluation results in Fig. 4 (b) of the main manuscript and Table 9.

Following prior CPath foundation models such as TITAN [20] and UNI [12], we trained a linear classifier on top of hierarchical features extracted from the Stage I MLLM-HWSI encoder. The classifier was optimized using an ℓ_2 -regularized L-BFGS solver from `scikit-learn`, with a maximum of 500 iterations. For datasets lacking a dedicated validation set, we used default settings with $\ell_2 = 1$ and 1,000 iterations to ensure stable convergence. The linear classifier was trained using cross-entropy loss on frozen embeddings aggregated across cell-, patch-, region-, and slide-level tokens.

Table 9 presents results across six public datasets, comparing MLLM-HWSI to leading CPath foundation models, including TITAN, FOCUS, GigaPath, and UNI. O MLLM-HWSI model consistently achieves the best performance across all datasets and metrics, attaining the highest F1-score (F) and balanced accuracy (BA) on PANDA (0.882 / 0.867), EBRAINS30 (0.833 / 0.803), BRACS (0.603 / 0.571), UBC-Ocean (0.968 / 0.961), TCGA-OT (0.789 / 0.766), and IMP-CRC (0.951 / 0.981). These substantial improvements over strong baselines such as TITAN (0.836 / 0.823 on PANDA) and UNI (0.809 / 0.757 on PANDA) demonstrate that hierarchical vision–language alignment yields highly discriminative and transferable WSI representations. Overall, the linear probe results confirm that MLLM-HWSI learns semantically structured, multi-scale embeddings that generalize effectively across organs, cancer types, and dataset domains—validating the effectiveness of hierarchical pre-training in capturing biologically meaningful and diagnostic features.

Variants	Semantic Patch Filtering		PANDA (BA)	EBRAINS (BA)	WSI-VQA (A)	SlideBench-VQA (BCNB)(A)
	HPS	DPS				
a. MLLM-HWSI	✓	✓	0.748	0.612	0.692	0.687
b. MLLM-HWSI	✓	×	0.731	0.592	0.674	0.663
c. MLLM-HWSI	×	✓	<u>0.741</u>	<u>0.606</u>	<u>0.684</u>	<u>0.676</u>
d. MLLM-HWSI	×	×	0.711	0.566	0.664	0.657
Variants	Semantic Patch Filtering		PANDA (BA)	EBRAINS (BA)	WSI-VQA (A)	SlideBench-VQA (BCNB)(A)
	K-means	DPS				
e. MLLM-HWSI	✓	✓	0.702	0.571	0.661	0.654
f. MLLM-HWSI	✓	×	0.683	0.554	0.641	0.644

Table 6. **Effect of Semantic Patch Filtering.** Comparison of different combinations of Heterogeneous Patch Selection (HPS), Diagnostically Relevant Patch Selection (DPS), and K-means clustering in MLLM-HWSI. The combination of HPS and DPS yields the best overall accuracy, highlighting their complementary roles in selecting diverse and diagnostic patches.

top- <i>k</i> Value	PANDA [7] (BA)	EBRAINS [42] (BA)	WSI-VQA [10] (A)	SlideBench-VQA (BCNB) [13](A)
16	0.711	0.588	0.676	0.641
32	0.731	0.605	0.681	0.663
48	0.748	0.612	0.692	0.687
64	<u>0.743</u>	0.604	<u>0.686</u>	<u>0.682</u>
96	0.740	0.608	0.685	0.678
128	0.735	<u>0.610</u>	0.684	0.675

Table 7. **Influence of top-*k* in ViT_{cell-cell}.** Performance with different top-*k* values in the Diagnostically Relevant Patch Selection (DPS) module. The best results are achieved at top-*k* = 48, indicating optimal diagnostic coverage and compactness.

6. WSI-Level Report Generation Qualitative Results (Tables 10-14)

We conducted an extensive qualitative comparison of pathology report generation to evaluate the interpretive and diagnostic reasoning capabilities of MLLM-HWSI against SOTA CPath models, including WSI-LLaVA, MI-Gen, Hist-Gen, Quilt-LLaVA, and GPT-4o. Tables 10–14 illustrate representative examples covering multiple diagnostic contexts—morphological description, global architecture analysis, key diagnostic feature identification, molecular subtyping, and TNM staging.

Across all examples, MLLM-HWSI produces reports that are nearly indistinguishable from expert-authored ground truth, demonstrating close semantic and morphological alignment. Its outputs consistently capture fine-grained histological detail—including nuclear pleomorphism, keratinization, intercellular bridges, and mitotic figures—while preserving global structural context, such as tumor organization and invasion patterns. The generated descriptions are linguistically coherent, clinically interpretable, and free from redundant or hallucinated content that often appears in baseline models.

In morphological and global description tasks (Tables 10–11), MLLM-HWSI accurately describes both cellular morphology and tissue-level architecture, surpassing prior models that either miss key features or overgeneralize findings. For diagnostic and molecular interpretation (Tables 12–13), the model correctly identifies defining histologic and molecular attributes, such as papillary architecture, psammoma bodies, and HPV-negative subtypes, aligning precisely with ground-truth annotations. In the staging example (Table 14), MLLM-HWSI achieves perfect correspondence with clinical staging guidelines, correctly reporting T3 N2 M0 without deviation.

Overall, these qualitative analyses highlight that MLLM-HWSI not only surpasses all competing models in accuracy and language fluency but also demonstrates clinically grounded, evidence-based reasoning. By aligning hierarchical WSI features with pathology-specific language, MLLM-HWSI generates diagnostic narratives that faithfully replicate expert interpretation—bridging the gap between automated analysis and human-level pathological reporting.

7. Pre-training Details of MLLM-HWSI

The pre-training of MLLM-HWSI is organized into three sequential stages: (i) hierarchical WSI–text alignment, (ii) hierarchical feature-space alignment, and (iii) task-specific instruction tuning. Stages I and II utilize 9,642 WSI–caption pairs from the WSIBench dataset [34] covering diverse cancer types, while Stage III employs 175,450 WSI-level VQA pairs from the same source for instruction fine-tuning.

Overall, the training process is divided into three stages, i.e., hierarchical WSI–text alignment, hierarchical feature space alignment, and task-specific instruction tuning. In stage I and II, we used 9,642 WSIs–caption pairs from the WSIBench dataset [34]. In stage III, we used 175,450 WSI-level VQA pairs from the WSIBench dataset [34].

Table 8. WSI-level Zero-shot classification performance comparison results with SOTA CPath models across six datasets.

Method	PANDA		EBRAINS		BRACS		UBC-Ocean		TCGA-OT		IMP-CRC	
	F	BA	F	BA	F	BA	F	BA	F	BA	F	BA
PLIP	0.288	0.235	0.013	0.080	0.214	0.203	0.376	0.345	0.203	0.223	0.523	0.655
PathCLIP	0.461	0.455	0.223	0.187	0.281	0.309	0.657	0.612	0.304	0.334	0.560	0.700
CPLIP	0.445	0.420	0.253	0.233	0.294	0.288	0.706	0.653	0.431	0.405	0.591	0.733
CONCH	0.596	0.566	0.304	0.278	0.336	0.344	0.786	0.807	0.488	0.532	0.637	0.833
QuiltNet	0.532	0.509	0.229	0.201	0.321	0.312	0.753	0.776	0.486	0.506	0.608	0.788
MR-PLIP	<u>0.701</u>	0.681	0.332	0.314	<u>0.403</u>	<u>0.411</u>	0.855	0.833	0.506	0.541	<u>0.679</u>	0.809
SlideChat	0.605	0.633	0.479	0.326	0.248	0.255	0.861	<u>0.902</u>	0.493	0.487	0.648	0.809
PathGenCLIP	0.511	0.488	0.255	0.221	0.295	0.288	0.786	0.756	0.498	0.522	0.612	0.723
MI-Zero	0.405	0.386	0.253	0.233	0.261	0.241	0.807	0.786	0.506	0.486	0.585	0.666
KEP	0.476	0.455	0.209	0.193	0.244	0.221	0.734	0.721	0.446	0.456	0.598	0.687
TITAN	0.621	0.608	<u>0.365</u>	<u>0.543</u>	0.385	0.381	0.908	0.865	<u>0.713</u>	<u>0.616</u>	<u>0.723</u>	<u>0.861</u>
WSI-LLaVA	0.668	0.644	0.389	0.501	0.294	0.289	0.872	0.881	<u>0.517</u>	<u>0.523</u>	0.692	0.823
PRISM	0.544	0.521	0.263	0.279	0.322	0.334	0.753	0.765	0.538	0.460	0.624	0.743
Proposed MLLM-HWSI	0.722	0.748	0.458	0.612	0.446	0.456	0.952	0.922	0.748	0.666	0.767	0.908

Table 9. WSI-level classification results and comparisons using linear probe evaluation and weakly supervised MIL-based classification with SOTA CPath models across six datasets.

Method	PANDA		EBRAINS30		BRACS		UBC-Ocean		TCGA-OT		IMP-CRC	
	F	BA	F	BA	F	BA	F	BA	F	BA	F	BA
HIPT	0.687	0.654	0.702	0.677	0.334	0.288	0.766	0.706	0.512	0.488	0.718	0.801
CTransPath	0.752	0.691	0.597	0.514	0.398	0.355	0.788	0.733	0.566	0.544	0.749	0.833
REMEDIS	0.766	0.711	0.471	0.382	0.367	0.331	0.733	0.706	0.504	0.455	0.772	0.843
CHIEF	0.745	0.724	0.706	0.688	0.413	0.387	0.823	0.789	0.640	0.528	0.701	0.781
DinoPath	0.682	0.706	0.771	<u>0.755</u>	0.394	0.361	0.844	0.821	0.586	0.556	0.792	0.855
Virchow	0.741	0.728	0.723	0.701	0.459	0.422	0.902	0.889	0.656	0.564	0.708	0.755
RudolfV	0.653	0.677	0.706	0.688	0.438	0.401	0.881	0.865	0.607	0.596	0.732	0.786
UNI	0.809	0.757	0.746	0.675	<u>0.538</u>	<u>0.504</u>	0.940	0.922	0.657	0.633	0.814	0.881
GigaPath	0.789	0.794	0.704	0.687	0.507	0.477	0.901	0.889	0.659	0.543	0.791	0.856
TITAN	<u>0.836</u>	<u>0.823</u>	<u>0.786</u>	0.735	0.511	0.400	<u>0.956</u>	<u>0.933</u>	<u>0.764</u>	<u>0.704</u>	<u>0.903</u>	<u>0.946</u>
FOCUS	0.804	0.782	0.733	0.671	0.474	0.451	0.903	0.841	0.685	0.605	0.829	0.877
Proposed MLLM-HWSI	0.882	0.867	0.833	0.803	0.603	0.571	0.968	0.961	0.789	0.766	0.951	0.981

Stage I (Hierarchical WSI-Text Alignment). In this stage, we align multi-scale WSI representations with their textual counterparts. The learning rate is set to 1×10^{-3} , and the batch size to 64. Only the two-layer projection matrices responsible for vision-language alignment are optimized, while both the hierarchical encoders and the text encoder remain frozen. The model is trained for 50 epochs with a temperature coefficient of 0.02 to regulate the contrastive learning objective.

Stage II (Hierarchical Feature-Space Alignment). During this phase, both the multi-scale visual encoder and the LLM remain frozen, and training focuses exclusively on refining the hierarchical projection layers to harmonize feature distributions across modalities. The learning rate is maintained at 1×10^{-3} , using a global batch size of 256 for one epoch. The maximum input length is set to 2048 tokens, with no weight decay and a warmup ratio of 0.03 to ensure stable optimization.

Stage III (Instruction Fine-Tuning). This stage enables multimodal reasoning by tuning the LLM jointly with the hierarchical projection layers while keeping the hierarchical encoder frozen. The learning rate is reduced to 2×10^{-5} , with a global batch size of 128 and a maximum sequence length of 2048. Weight decay remains 0, and the warmup ratio is fixed at 0.03. To achieve parameter-efficient adaptation, we apply LoRA (Low-Rank Adaptation) with a rank of 128 and $\alpha = 256$. Training is performed using DeepSpeed ZeRO-3 for distributed optimization and BF16 precision with TensorFloat32 acceleration, improving computational efficiency while maintaining numerical stability.

8. Computational Pathology Datasets

To comprehensively evaluate MLLM-HWSI across a diverse range of CPath tasks, we employed multiple publicly available WSI datasets spanning classification, visual question answering (VQA), report generation, retrieval, and cap-

tioning benchmarks.

For **WSI classification**, including both zero-shot and linear probe evaluations, we used six standard benchmarks: BRACS [5], PANDA [7], IMP-CRC [39], TCGA-OT [20, 38], EBRAINS [42], and UBC-Ocean [3]. These datasets encompass a wide spectrum of organs, cancer subtypes, and histological grading systems, ensuring robust cross-domain generalization.

For the zero-shot **VQA** task, we adopted four multi-modal benchmarks: WSI-Bench (4,119 pairs) [34], WSI-VQA (8,672 pairs) [10], SlideBench-VQA (BCNB) (7,247 pairs) [13], and SlideBench-VQA (TCGA) (7,824 pairs) [13]. Together, these datasets evaluate the model’s ability to reason over morphological, diagnostic, and clinical questions at the slide level.

For **report generation**, we used the WSI-Bench (208 WSI-report pairs) [35] and HistGen (700 pairs) [25] datasets, both curated to assess automatic report synthesis grounded in morphological evidence.

For the **WSI retrieval** task, we evaluated on TCGA-OT [20, 38], EBRAINS [42], and IMP-CRC [39], enabling assessment of large-scale visual similarity retrieval in diagnostic contexts.

For **cross-modal retrieval**, we utilized the TCGA Reports dataset [20, 46], which links WSIs with associated clinical and textual records to evaluate bidirectional alignment between visual and textual representations.

Finally, for **caption generation**, we used the SlideBench dataset [13], designed for producing concise, pathology-grounded descriptions of WSIs.

Collectively, these datasets provide a comprehensive evaluation suite for assessing MLLM-HWSI’s performance across diagnostic interpretation, reasoning, and language grounding tasks in computational pathology.

1. BRACS (7 classes) [5] consists of 547 H&E FFPE WSIs of breast tumors (benign, atypical, and malignant) collected from 189 patients. The cases are annotated at two levels: a coarse-grained level of three classes (benign tumors: 265, atypical tumors: 89, malignant tumors: 193) and a fine-grained level of seven subtypes (including invasive carcinoma, ductal carcinoma in situ, and various benign/atypical hyperplasias). The dataset is divided into five label-stratified, patient-level splits using a 60:20:20 ratio (approx. 302:94:151 slides) for training, validation, and testing.

2. UBC-Ocean (5 Classes) [3] comprises 538 WSIs, with 527 meeting foreground tissue criteria, for ovarian cancer subtyping. The dataset covers five distinct subtypes: Clear Cell (CC), Endometrioid (EC), High-Grade Serous Carcinoma (HGSC), Low-Grade Serous Carcinoma (LGSC), and Mucinous Carcinoma (MC). The dataset is divided in a stratified fashion into train:validation:test sets with

approximately 369:52:106 WSIs, respectively.

3. TCGA-OT (46 Classes) [20, 38] is a pan-cancer subtyping dataset derived from TCGA, consisting of 11,186 H&E FFPE diagnostic histopathology WSIs of primary tumors. All WSIs are classified into 46 distinct cancer types based on the OncoTree classification system, with each class represented by at least 50 samples. Slides were rigorously curated by excluding frozen tissues, metastatic/recurrent tumors, and slides lacking magnification or tumor tissue. The dataset is split into training, validation, and test folds of 8,226:1,612:1,348 samples, respectively, while ensuring all slides from the same source site remain within a single split.

4. EBRAINS (30 classes) dataset [42] features H&E-stained whole-slide images (WSIs) of brain tissue sourced from The Digital Brain Tumour Atlas. For our study, we utilized a subset of 2,319 WSIs (out of 3,114 total), mirroring the selection process used for the CONCH dataset [36]. This defined a 30-class fine-grained brain tumor subtyping task, including only diagnostic labels with at least 30 slides. We established the WSI counts per class to match those in CONCH. For the supervised task, the 2,319 slides were split 50%-25%-25% into training (1,151 slides), validation (595 slides), and testing (573 slides). This 573-slide testing split was also used as the zero-shot test set.

5. PANDA (6 classes) is the International Society of Urological Pathology (ISUP) grading task derived from the PANDA challenge [7]. This dataset comprises prostate cancer core needle biopsies. We utilized a subset of 9,555 Whole Slide Images (WSIs) after excluding noisy labels from the original 10,616 slides. These 9,555 slides are distributed across the six ISUP grades as follows: Grade 0 (2,603), Grade 1 (2,399), Grade 2 (1,209), Grade 3 (1,118), Grade 4 (1,124), and Grade 5 (1,102). For experiments, the dataset was partitioned into standard 80% training, 10% validation, and 10% test sets (7,647:954:954 WSIs).

6. IMP-CRC (3 Classes) [39] is a colorectal cancer dataset containing 5,333 H&E FFPE biopsy and polypectomy WSIs from the IMP Diagnostics laboratory. Cases are classified into three distinct categories: Non-neoplastic (847 slides), Low-grade lesions (2847 slides) which include conventional adenomas with low-grade dysplasia, and High-grade lesions (1639 slides) encompassing conventional adenomas with high-grade dysplasia, intramucosal carcinomas, and invasive adenocarcinomas. The dataset is label-stratified and split into train:validation:test sets using a 60:20:20 ratio, resulting in 3546:887:900 slides, respectively.

7. WSI-Bench [34] is a large-scale VQA dataset specifically designed for WSIs. It contains a total of 179,569 VQA pairs. The training set comprises 175,450 pairs across 9,642 WSIs (122,133 open-ended and 53,317 closed-ended questions). The test set consists of 4,119 VQA pairs from 208 WSIs (2,838 open-ended and 1,281 closed-ended questions). Additionally, a specific subset of 208 VQA pairs is dedicated to report generation.

8. WSI-VQA dataset [9] contains 977 whole-slide images (WSIs), which are paired with a total of 8,672 question-and-answer (QA) pairs. On average, this amounts to approximately 8.9 QA pairs per WSI. The QA pairs are composed of 4,535 close-ended questions and 4,137 open-ended questions.

9. SlideBench-VQA (BNCB) [13] is a dataset comprising 7,247 Visual Question Answering (VQA) pairs derived from 1,058 patients. Its primary purpose is to evaluate the zero-shot generalization capability of models like SlideChat across seven distinct classification tasks.

10. SlideBench-VQA (TCGA) [13] is a VQA dataset specifically focused on WSIs sourced from The Cancer Genome Atlas (TCGA). The dataset comprises 7,827 VQA pairs, which cover 13 distinct WSI categories. The 2451 overlapping samples of SlideBench-VQA (test split) with WSI-Bench were not used during training. All evaluations were performed on held-out test splits. Our zero-shot results, therefore, reflect generalization to unseen WSIs.

11. HistGen-Report [25] is a WSI dataset designed for report generation. It comprises 7,753 WSI-report pairs sourced from the TCGA platform. The diagnostic reports were subsequently refined using large language models to ensure high quality, coherence, and diagnostic relevance.

12. TCGA-Reports [20, 46] is a dataset containing pathology reports sourced from The Cancer Genome Atlas (TCGA) data portal. The dataset was compiled from 11,108 pathology report PDFs, corresponding to 11,010 patients.

References

- [1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 4, 5
- [2] Alexi Baidoshvili, Mariam Khacheishvili, Jeroen AWM van der Laak, and Paul J van Diest. A whole-slide imaging based workflow reduces the reading time of pathologists. *Pathology International*, 73(3):127–134, 2023. 2
- [3] Ali Bashashati, Hossein Farahani, OTTA Consortium, Anthony Karnezis, Ardalan Akbari, Sirim Kim, Ashley Chow, Sohier Dane, Allen Zhang, and Maryam Asadi. Ubc ovarian cancer subtype classification and outlier detection (ubcocean), 2023. 8
- [4] Samar Betmouni. Diagnostic digital pathology implementation: learning from the digital health experience. *Digital Health*, 7:20552076211020240, 2021. 1
- [5] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubiarta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093, 2022. 8
- [6] Romain Brixel, Sebastien Bougleux, Olivier Lézoray, Yann Caillot, Benoit Lemoine, Mathieu Fontaine, Dalal Nebati, and Arnaud Renouf. Whole slide image quality in digital pathology: review and perspectives. *IEEE Access*, 10: 131005–131035, 2022. 1, 2
- [7] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F. Steiner, Hester van Boven, Robert Vink, Christina Hulsbergen-van de Kaa, Jeroen van der Laak, Mahul B. Amin, Andrew J. Evans, Theodorus van der Kwast, Robert Allan, Peter A. Humphrey, Henrik Grönberg, Hemamali Samaratunga, and ... the PANDA challenge consortium. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature Medicine*, 28: 154–163, 2022. 2, 3, 6, 8
- [8] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 4, 5
- [9] Pingyi Chen, Chenglu Zhu, Sunyi Zheng, Honglin Li, and Lin Yang. Wsi-vqa: Interpreting whole slide images by generative visual question answering. In *European Conference on Computer Vision (ECCV) 2024*, pages 401–417, 2025. 9
- [10] Pingyi Chen, Chenglu Zhu, Sunyi Zheng, Honglin Li, and Lin Yang. Wsi-vqa: Interpreting whole slide images by generative visual question answering. In *European Conference on Computer Vision*, pages 401–417. Springer, 2025. 2, 3, 6, 8
- [11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1, 2
- [12] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. 5
- [13] 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 *Proceedings of the*

- Computer Vision and Pattern Recognition Conference*, pages 5134–5143, 2025. 1, 2, 3, 6, 8, 9
- [14] Norman F Cheville. *Cell pathology*. Number Edition 2. 1983. 1
- [15] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. 4, 5
- [16] Toby C Cornish, Ryan E Swapp, and Keith J Kaplan. Whole-slide imaging: routine pathologic diagnosis. *Advances in anatomic pathology*, 19(3):152–159, 2012. 1
- [17] Miao Cui and David Y Zhang. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4): 412–422, 2021. 1
- [18] Charles Frederick Albert Culling, RT Allison, and WT Barr. *Cellular pathology technique*. Elsevier, 2014. 1
- [19] Neofytos Dimitriou, Ognjen Arandjelović, and David J Harrison. Magnifying networks for histopathological images with billions of pixels. *Diagnostics*, 14(5):524, 2024. 1
- [20] Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F. K. Williamson, Bowen Chen, Cristina Almagro-Perez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Daisuke Komura, Akihiro Kawabe, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. Multimodal whole slide foundation model for pathology, 2024. 5, 8, 9
- [21] Filippo Frassetto, Salvatore Garozzo, Gian Franco Zannoni, Liron Pantanowitz, and Esther Diana Rossi. Routine digital pathology workflow: the catania experience. *Journal of pathology informatics*, 8(1):51, 2017. 1, 2
- [22] Thomas J Fuchs and Joachim M Buhmann. Computational pathology: challenges and promises for tissue analysis. *Computerized Medical Imaging and Graphics*, 35(7-8): 515–530, 2011. 1
- [23] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019. 2
- [24] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 4, 5
- [25] Zhengrui Guo, Jiabo Ma, Yingxue Xu, Yihui Wang, Liansheng Wang, and Hao Chen. Histgen: Histopathology report generation via local-global feature encoding and cross-modal context interaction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 189–199. Springer, 2024. 8, 9
- [26] Matthew G Hanna, Anil Parwani, and Sahussapont Joseph Sirintrapun. Whole slide imaging: technology and applications. *Advances in anatomic pathology*, 27(4):251–259, 2020. 1
- [27] Shuko Harada and Diana Morlote. Molecular pathology of colorectal cancer. *Advances in anatomic pathology*, 27(1): 20–26, 2020. 2
- [28] Assia Hijazi, Carlo Bifulco, Pamela Baldin, and Jérôme Galon. Digital pathology for better clinical practice. *Cancers*, 16(9):1686, 2024. 2
- [29] Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, et al. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024. 2, 4
- [30] Carolyn Hutter and Jean Claude Zenklusen. The cancer genome atlas: creating lasting value beyond its data. *Cell*, 173(2):283–285, 2018. 1
- [31] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36, 2024. 5
- [32] Nfn Kiran, FNU Sapna, FNU Kiran, Deepak Kumar, FNU Raja, Sheena Shiwlani, Antonella Paladini, FNU Sonam, Ahmed Bendari, Raja Sandeep Perakash, et al. Digital pathology: transforming diagnosis in the digital age. *Cureus*, 15(9), 2023. 1
- [33] Bao Li, Zhenyu Liu, Song Zhang, Xiangyu Liu, Caixia Sun, Jiangang Liu, Bensheng Qiu, and Jie Tian. Nuhtc: A hybrid task cascade for nuclei instance segmentation and classification. *Medical Image Analysis*, 103:103595, 2025. 2
- [34] Yuci Liang, Xinheng Lyu, Meidan Ding, Wenting Chen, Jipeng Zhang, Yuexiang Ren, Xiangjian He, Song Wu, Sen Yang, Xiyue Wang, Xiaohan Xing, and Linlin Shen. Wsi-llava: A multimodal large language model for whole slide image. *arXiv preprint arXiv:2412.02141*, 2024. 6, 8, 9
- [35] Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, and Linlin Shen. Wsi-llava: A multimodal large language model for whole slide image, 2025. 1, 3, 8
- [36] Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pre-trained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19764–19775, 2023. 8
- [37] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024. 5
- [38] F. Mahmood et al. Tcga-ot: A 46-class whole slide image dataset for oncotree classification, 2024. Accessed: 2025-09-27. 8
- [39] Pedro C Neto, Diana Montezuma, Sara P Oliveira, Domingos Oliveira, João Fraga, Ana Monteiro, João Monteiro, Liliana Ribeiro, Sofia Gonçalves, Stefan Reinhard, et al. An interpretable machine learning system for colorectal cancer

diagnosis from pathology slides. *NPJ precision oncology*, 8 (1):56, 2024. [8](#)

- [40] Linhao Qu, Kexue Fu, Manning Wang, Zhijian Song, et al. The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. *Advances in Neural Information Processing Systems*, 36, 2024. [2](#)
- [41] Shan E Ahmed Raza, Linda Cheung, Muhammad Shaban, Simon Graham, David Epstein, Stella Pelengaris, Michael Khan, and Nasir M. Rajpoot. Micro-net: A unified model for segmentation of various objects in microscopy images. *Medical Image Analysis*, 52:160–173, 2019. [2](#)
- [42] Thomas Roetzer-Pejrimovsky, Anna-Christina Moser, Baran Atli, Clemens Christian Vogel, Petra A Mercea, Romana Prihoda, Ellen Gelpi, Christine Haberler, Romana Höftberger, Johannes A Hainfellner, et al. The digital brain tumour atlas, an open histopathology resource. *Scientific Data*, 9(1):55, 2022. [2](#), [3](#), [6](#), [8](#)
- [43] Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12): 930–949, 2023. [1](#), [2](#)
- [44] Sara C Steele. Vocabulary intervention: A national survey of school-based speech–language pathologists. *Communication Disorders Quarterly*, 41(3):151–161, 2020. [1](#)
- [45] Martin Weigert and Uwe Schmidt. Nuclei instance segmentation and classification in histopathology images with stardist. In *The IEEE International Symposium on Biomedical Imaging Challenges (ISBIC)*, 2022. [2](#)
- [46] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013. [8](#), [9](#)
- [47] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and Guanting Dong et al. Qwen2 technical report. Technical Report arXiv:2407.10671, CoRR, arXiv, 2024. [4](#), [5](#)
- [48] Mark D Zarella, Douglas Bowman, Famke Aeffner, Navid Farahani, Albert Xthona, Syeda Fatima Absar, Anil Parwani, Marilyn Bui, and Douglas J Hartman. A practical guide to whole slide imaging: a white paper from the digital pathology association. *Archives of pathology & laboratory medicine*, 143(2):222–234, 2019. [1](#)

Table 10. Qualitative comparison of pathology report generation across SOTA CPath models. The qualitative analysis illustrates how MLLM-HWSI produces reports that closely match expert-annotated ground truth, capturing both fine-grained cellular morphology and global architectural context. Compared with prior models (e.g., WSI-LLaVA, MI-Gen, Hist-Gen, Quilt-LLaVA, and GPT-4o), MLLM-HWSI generates linguistically coherent and diagnostically accurate descriptions that mirror expert reasoning, demonstrating superior grounding between visual evidence and clinical language. Green: matched ground-truth content; Red: deviations; Orange: ground truth content missing in model response; Underlined: template language.

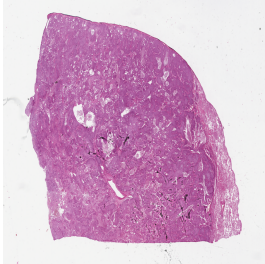
	
WSI	
Question	What are the main observations in the whole slide histopathology image, and what is the likely diagnosis?
WSI-LLaVA	Microscopic observation of the pathology slide reveals moderately differentiated squamous cell carcinoma characterized by nests and sheets of atypical squamous epithelial cells with intercellular bridges and keratinization. The tumor cells exhibit moderate nuclear pleomorphism with enlarged, hyperchromatic nuclei and prominent nucleoli. Mitotic figures are present, indicating active cell division . There is no evidence of vascular invasion, necrosis, or calcification within the examined tissue. Based on these observations, the diagnosis is moderately differentiated keratinizing squamous carcinoma.
MI-Gen	Microscopic observation of the pathology slide reveals a poorly differentiated squamous cell carcinoma characterized by a lack of squamous cell carcinoma. The tumor exhibits a lack of squamous differentiation with a lack of squamous differentiation. The tumor cells exhibit moderate nuclear pleomorphism with variability in nuclear size and shape and there is a moderate rate of cell division. Vascular invasion is not identified. Necrosis is not observed within the tumor. Based on these observations, the diagnosis is squamous cell carcinoma.
Hist-Gen	Microscopic observation of the pathology slide reveals a poorly differentiated squamous cell carcinoma characterized by irregular nests and sheets of atypical squamous cells with intercellular bridges and keratinization. The tumor cells exhibit a high degree of nuclear pleomorphism . There is no evidence of vascular invasion or necrosis within the tumor tissue. The surrounding lung parenchyma shows areas of chronic inflammation and chronic inflammation. Based on these observations, the diagnosis is squamous cell carcinoma, poorly differentiated .
Quilt-LLaVA	Microscopic observation of the pathology slide reveals sheets and nests of atypical squamous cells with moderate differentiation. There is evidence of focal keratinization, characterized by the presence of keratin pearls within the tumor tissue. Intercellular bridges are noted among the squamous cells, indicating differentiation towards squamous epithelium. The cellular morphology shows moderate pleomorphism, with variability in nuclear size and shape, and occasional prominent nucleoli. Mitotic figures are present but not excessively numerous. Vascular invasion is not specified in the observations, and necrosis is not prominently noted. Based on these observations, the diagnosis is squamous cell carcinoma, moderately differentiated.
GPT-4o	This histopathology image shows a well-circumscribed, nodular lesion with a dense, cellular stroma. There are areas of necrosis and hemorrhage, and the tissue architecture appears disrupted. The lesion has a heterogeneous appearance with varying cellular densities and some fibrotic regions. Diagnosis: The features are consistent with a uterine fibroid (leiomyoma), which is characterized by a smooth muscle cell growth forming a circumscribed mass with areas of degeneration.
MLLM-HWSI	<u>Microscopic observation of the pathology slide reveals sheets and nests of atypical squamous epithelial cells with moderate differentiation.</u> There is clear evidence of focal keratinization, highlighted by the formation of keratin pearls within the tumor tissue. Intercellular bridges are well identified among the tumor cells, confirming differentiation toward squamous epithelium. The nuclei display moderate pleomorphism with variation in size and shape and occasional prominent nucleoli. Mitotic figures are present but not numerous, indicating a moderate proliferative index. There is no evidence of vascular invasion, and necrosis is not apparent in the examined sections. <u>Based on these findings, the features are consistent with a moderately differentiated keratinizing squamous cell carcinoma.</u>
Ground Truth	Microscopic observation of the pathology slide reveals sheets and nests of atypical squamous cells with moderate differentiation. There is evidence of focal keratinization, characterized by the presence of keratin pearls within the tumor tissue. Intercellular bridges are noted among the squamous cells, indicating differentiation towards squamous epithelium. The cellular morphology shows moderate pleomorphism, with variability in nuclear size and shape, and occasional prominent nucleoli. Mitotic figures are present but not excessively numerous. Vascular invasion is not specified in the observations, and necrosis is not prominently noted. Based on these observations, the diagnosis is squamous cell carcinoma, moderately differentiated.

Table 11. A comparative example of global morphology description outputs from different CPath models. Green: matched ground-truth content; Red: deviations; Orange: ground truth content missing in model response; Underlined: template language.

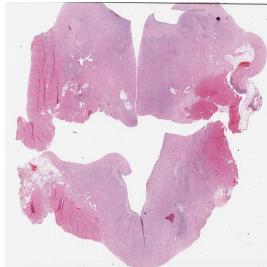
	
WSI	
Question	Can you provide an overall description of the whole slide image (WSI)?
WSI-LLaVA	The whole slide image reveals exuberant granulation tissue characterized by notable fibroblastic proliferation and myxoid changes. The cellular component is primarily composed of spindle-shaped cells within a myxoid matrix. There is an absence of normal tissue architecture as the granulation tissue dominates the observed section. Importantly, no residual sarcoma tissue is identifiable within this examined section.
Quilt-LLaVA	The whole slide image reveals a tumor with a diffuse architectural pattern, characterized by a mixture of alveolar formations and glandular cell differentiation. There is a notable presence of large cells with prominent nucleoli and moderate nuclear atypia. The mitotic activity is moderate, with 8 mitoses per 10 high power fields. The tumor shows significant vascular invasion, with tumor cells infiltrating the renal vein. Additionally, there is focal necrosis within the tumor tissue.
WSI-VQA	The whole slide image reveals a biphasic tumor composed of spindle-shaped cells arranged in a storiform pattern. The spindle cells are arranged in a storiform pattern. The spindle cells are arranged in a storiform pattern. There is a notable absence of necrosis or calcification within the tumor.
GPT-4o	The whole slide image appears to be a tissue section stained with hematoxylin and eosin (H&E), commonly used in histopathology. The sample is divided into several sections, likely representing different orientations or parts of the tissue. – *Coloration*: The pink and purple hues suggest typical H&E staining, where eosin stains cytoplasmic components pink and hematoxylin stains nuclear material blue or purple. – *Tissue Architecture*: The sections show varying thickness with folded tissue smooth and irregular contours, indicating possible different tissue types or regions. – *Other Observations*: There are areas that seem to contain more densely packed cells and some regions with more open or less cellular areas, suggesting different tissue densities or types. A detailed examination at higher magnification would be necessary for further analysis, such as identifying specific pathological features or tissues.
MLLM-HWSI	The WSI demonstrates exuberant granulation tissue with extensive fibroblastic proliferation embedded in a myxoid stroma. The spindle-shaped fibroblasts are loosely arranged within the extracellular matrix, accompanied by delicate capillary networks and scattered inflammatory infiltrates. No evidence of residual malignant or sarcomatous components is observed. The overall architecture is dominated by reactive granulation tissue replacing normal parenchymal organization, consistent with post-treatment reparative changes.
Ground Truth	The whole slide image demonstrates exuberant granulation tissue characterized by notable fibroblastic proliferation and myxoid changes. The cellular component is primarily composed of spindle-shaped cells within a myxoid matrix. There is an absence of normal tissue architecture as the granulation tissue dominates the observed section. Importantly, no residual sarcoma tissue is identifiable within this examined section.

Table 12. A comparative example of key diagnostic description outputs from different CPath models. Green: matched ground-truth content; Red: deviations; Orange: ground truth content missing in model response; Underlined: template language.

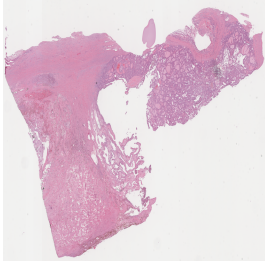
	
WSI	
Question	Which observed features are most significant for making a diagnosis?
WSI-LLaVA	The most significant features include the papillary architecture with fibrovascular cores, the presence of psammoma bodies, and the nuclear characteristics such as overlapping, grooves, and intranuclear cytoplasmic inclusions. The absence of vascular invasion, necrosis, and calcification also plays a crucial role in the diagnostic process.
Quilt-LLaVA	The most significant features include the irregular clusters of epithelial cells with poor glandular differentiation, marked nuclear pleomorphism, and the presence of perineural invasion. The high mitotic activity and evidence of lymphovascular invasion are also critical, as they indicate aggressive tumor behavior.
WSI-VQA	The most significant features include the papillary architecture with fibrovascular cores and the presence of fibrovascular cores. The absence of vascular invasion and necrosis further supports the diagnosis. The absence of necrosis and calcification further supports the diagnosis.
GPT-4o	The whole slide image appears to be a tissue section stained with hematoxylin and eosin (H&E), commonly used in histopathology. The sample is divided into several sections, likely representing different orientations or parts of the tissue. – *Coloration*: The pink and purple hues suggest typical H&E staining, where eosin stains cytoplasmic components pink and hematoxylin stains nuclear material blue or purple. – *Tissue Architecture*: The sections show varying thickness with folded tissue smooth and irregular contours, indicating possible different tissue types or regions. – *Other Observations*: There are areas that seem to contain more densely packed cells and some regions with more open or less cellular areas, suggesting different tissue densities or types. A detailed examination at higher magnification would be necessary for further analysis, such as identifying specific pathological features or tissues.
MLLM-HWSI	The diagnostic features most critical for interpretation include a prominent papillary architecture with well-formed fibrovascular cores lined by epithelial cells showing characteristic nuclear overlapping, grooves, and intranuclear cytoplasmic inclusions. Multiple psammoma bodies are present within the papillary stalks. There is no evidence of vascular, lymphatic, or perineural invasion, and no areas of necrosis are identified. These cytologic and architectural features are diagnostic of a papillary-patterned neoplasm consistent with papillary carcinoma morphology.
Ground Truth	The whole slide image demonstrates exuberant granulation tissue characterized by notable fibroblastic proliferation and myxoid changes. The cellular component is primarily composed of spindle-shaped cells within a myxoid matrix. There is an absence of normal tissue architecture as the granulation tissue dominates the observed section. Importantly, no residual sarcoma tissue is identifiable within this examined section.

Table 13. A comparative example of molecular subtyping outputs from different CPath models.

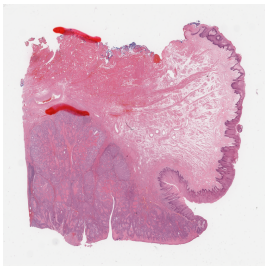
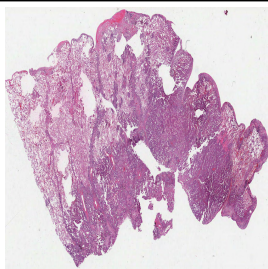
	
WSI	
Question	Based on the IHC results showing P16 negativity and HR, HPV ISH status negativity, what is the molecular subtype in this head and neck squamous cell carcinoma?
WSI-LLaVA	The molecular subtype is HNSC HPV .
Quilt-LLaVA	The molecular subtype is HNSC HPV .
WSI-VQA	the molecular subtype is hnscc unk .
GPT-4o	The molecular subtype of head and neck squamous cell carcinoma with P16 negativity and HR HPV ISH negativity is typically HPV-negative. This subtype is generally associated with non-viral risk factors such as smoking and alcohol use.
MLLM-HWSI	Given the P16 negativity and HR HPV ISH negativity, the molecular subtype is correctly classified as HNSC HPV-negative. This reflects a non-viral, keratinizing phenotype often linked to tobacco and alcohol-related carcinogenesis. These molecular features are characteristic of the HPV-independent pathway in head and neck squamous cell carcinoma.
Ground Truth	The molecular subtype is HNSC HPV

Table 14. A comparative example of staging outputs from different CPath models.

<p>WSI</p>	
<p>Question</p>	<p>What is the TNM staging for a patient with a 7 cm tumor, several affected lymph nodes, but no distant metastasis?</p>
<p>WSI-LLaVA</p>	<p>The TNM staging for this patient is T3 N2 M0.</p>
<p>Quilt-LLaVA</p>	<p>The TNM staging for this patient is T3 N2 M0.</p>
<p>WSI-VQA</p>	<p>the tnm staging for this patient is t2 n2 m0.</p>
<p>GPT-4o</p>	<p>To determine the TNM staging, we need to consider the following components:- T (Tumor size): The tumor is 7 cm. This generally corresponds to T2 or T3, depending on the specific cancer type, but often T3 for many cancers. N(Nodes): "Several affected lymph nodes" typically indicates N1 or N2, depending on the exact number and extent of nodal involvement.- M (Metastasis): No distant metastasis is indicated as M0. Without specific details about the type of cancer, a general staging based on the information given could be T3N1M0 or T3N2M0. However, the exact stage can vary based on cancer type and guidelines, so consultation with specific cancer staging resources is advised for precise staging.</p>
<p>MLLM-HWSI</p>	<p>According to TNM criteria, the 7 cm primary tumor meets the definition of T3 due to its size exceeding 5 cm. The description of "several affected lymph nodes" corresponds to N2, indicating regional nodal spread to multiple ipsilateral nodes. No evidence of distant metastasis defines M0. Therefore, the overall TNM staging for this patient is accurately classified as T3 N2 M0.</p>
<p>Ground Truth</p>	<p>TNM staging for this patient is T3 N2 M0.</p>