# WSI-Agents: A Collaborative Multi-Agent System for Multi-Modal Whole Slide Image Analysis

Xinheng Lyu[1,2], Yuci Liang[3], Wenting Chen[4†], Meidan Ding[3], Jiaqi Yang[2,3], Guolin Huang[3,5], Daokun Zhang[2], Xiangjian He[2†], and Linlin Shen[1†]

[1] College of Artificial Intelligence, Shenzhen University, China
[2] School of Computer Science, University of Nottingham Ningbo China, China
[3] College of Computer Science and Software Engineering, Shenzhen University, China
[4] Department of Electrical Engineering, City University of Hong Kong, Hong Kong
[5] Wuyi University, China

**Abstract.** Whole slide images (WSIs) are vital in digital pathology, enabling gigapixel tissue analysis across various pathological tasks. While recent advancements in multi-modal large language models (MLLMs) allow multi-task WSI analysis through natural language, they often underperform compared to task-specific models. Collaborative multi-agent systems have emerged as a promising solution to balance versatility and accuracy in healthcare, yet their potential remains underexplored in pathology-specific domains. To address these issues, we propose WSI-Agents, a novel collaborative multi-agent system for multi-modal WSI analysis. WSI-Agents integrates specialized functional agents with robust task allocation and verification mechanisms to enhance both task-specific accuracy and multi-task versatility through three components: (1) a task allocation module assigning tasks to expert agents using a model zoo of patch and WSI level MLLMs, (2) a verification mechanism ensuring accuracy through internal consistency checks and external validation using pathology knowledge bases and domain-specific models, and (3) a summary module synthesizing the final summary with visual interpretation maps. Extensive experiments on multi-modal WSI benchmarks show WSI-Agents's superiority to current WSI MLLMs and medical agent frameworks across diverse tasks. Source code is available at https://github.com/XinhengLyu/WSI-Agents.

**Keywords:** Whole Slide Image · Multi-agent System · Digital Pathology

## 1 Introduction

Whole slide images (WSIs) are essential in digital pathology, providing gigapixel-scale digitized tissue samples. WSI analysis has evolved from early patch-based models targeting specific tasks (cancer classification [12, 13], grading [3, 2], and

[†]Corresponding authors: Wenting Chen (wentichen7-c@my.cityu.edu.hk), Linlin Shen (llshen@szu.edu.cn), and Xiangjian He (Sean.He@nottingham.edu.cn)

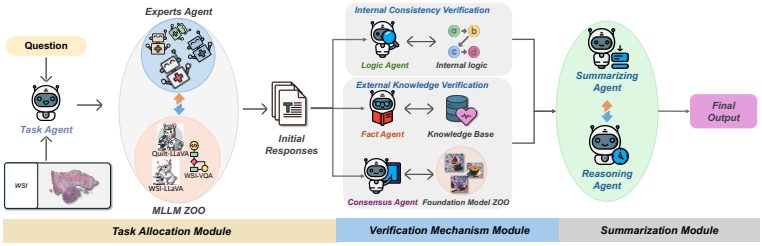



Fig. 1: The workflow of WSI-Agents with three main modules.

tumor detection [13, 7]) to advanced approaches handling gigapixel-scale analysis (survival prediction [30, 11], diagnosis [8, 32], report generation [10, 4], and visual question answering [5]). Despite impressive performance, most methods remain task-specific, limiting versatility. Consequently, there is a pressing need for comprehensive models capable of addressing multiple pathological tasks and adapting to diverse scenarios.

Recent advances in multi-modal large language models (MLLMs)[17, 6] have enabled WSI-level MLLMs[17, 6] capable of performing diverse pathological tasks through natural language interaction. Models such as WSI-LLaVA [17] and SlideChat [6] handle various tasks including morphological analysis, diagnosis, treatment planning, and report generation. While offering multi-task capability within a single framework, they exhibit significant performance gaps compared to specialized foundation models. For example, while dedicated foundation models achieve over 90% accuracy in cancer subtype classification, current MLLMs typically underperform by 15-30% on comparable diagnostic tasks [17, 6, 8, 27]. Therefore, improving task-specific performance while maintaining multi-task capabilities remains a critical challenge.

Recent AI agents [19, 16, 25, 31] can leverage multiple specialized models and real-world APIs, offering a potential solution to the accuracy-versatility trade-off. In medical applications, some agents [14, 29, 18] employ collaborative multi-agent systems for clinical reasoning, while others [9, 15] integrate external machine learning tools to enhance diagnostic accuracy. However, these approaches often either specialize too narrowly (e.g., chest X-rays) [9] or spread too broadly across imaging modalities [15], diluting domain expertise. Notably, these agents lack support for gigapixel WSIs, which are essential for comprehensive histopathological analysis. This limitation, coupled with insufficient pathology-specific knowledge, explains why general medical agents achieve only 40-50% accuracy on pathology-specific tasks compared to 75-80% on general medical tasks [14, 29]. Thus, incorporating both specialized pathology domain knowledge and robust WSI processing capabilities into medical agents is highly recommended for effective pathology analysis.

To address these challenges, we propose **WSI-Agents**, a novel collaborative multi-agent system for multi-modal WSI analysis that integrates specialized functional agents with robust verification mechanisms. As shown in Fig.1, WSI-Agentsconsists of three primary components: a **task allocation mod-**

**ule**, **verification mechanisms**, and a **summary module**. The task allocation module employs a **task agent** to interpret input tasks and assign them to specialized expert agents (morphology, diagnosis, report generation, etc.) who automatically select relevant models from our comprehensive model zoo, generating multiple initial responses.To ensure clinical accuracy, our **verification mechanism** includes **internal consistency verification** and **external knowledge verification**. Internal consistency verification uses a *logic agent* to check for contradictions and evidence validity. More critically, external knowledge verification addresses the *domain knowledge challenge* through fact and consensus agents. The *fact agent* validates responses against pathology knowledge bases constructed from medical literature, while the **consensus agent** leverages domain knowledge from WSI foundation models pre-trained on histopathological datasets. Each verification agent generates scores and maintains detailed logs in the memory module. Finally, the *summarizing agent* evaluates verification scores, selects optimal responses, and coordinates with *reasoning agents* to produce the final output. For visual interpretation, we integrate attention maps from multiple patch-level WSI models into a comprehensive visual interpretation map. Extensive experiments on two major multi-modal WSI benchmarks demonstrate WSI-Agents's superior performance compared to existing WSI MLLMs and medical agent frameworks.

## 2 WSI-Agents

**Overview.** The workflow of WSI-Agents is illustrated in Fig. 2, focusing on collaborative multi-agent WSI analysis. Given input WSI $I$ and question $x$, our system processes them through three modules: task allocation module, verification mechanism, and summary module. In the task allocation module, a task agent delegates responsibilities to specialized expert agents $A$ who select $M$ relevant models from our model zoo, generating multiple initial responses $\{y_i\}_{i=1}^{M}$. For verification, a logic agent extracts claims $\{C_i\}_{i=1}^{N}$ and evidences $\{E_i\}_{i=1}^{N}$, computes a compatibility matrix $G$ to represent the compatibility among claims, and evaluates evidence validity $\{V_i\}_{i=1}^{N}$ to determine internal consistency score $\phi_l$. The fact agent validates claims against a pathology knowledge base $B$ to compute verification score $\phi_f$. A consensus agent extracts diagnostic keywords $D$, compares them with foundation model results $\{R_i\}_{i=1}^{H}$ to compute MLLM-classifier agreement $\phi_a$, calculates inter-classifier agreement $\phi_b$ as a confidence factor, and multiplies these for classifier verification score $\phi_c$. In the summary module, a summarizing agent evaluates these verification scores and coordinates with reasoning agents to produce the final output $y$. For visualization, we integrate attention maps $\{m_i\}_{i=1}^{K}$ from $K$ patch-level models into a comprehensive interpretation map $m$, providing explainable visualization.

### 2.1 Task Allocation Module (TAM)

To adapt to diverse pathological scenarios, we introduce a task allocation module for WSI. This module comprises a task agent that interprets requirements and

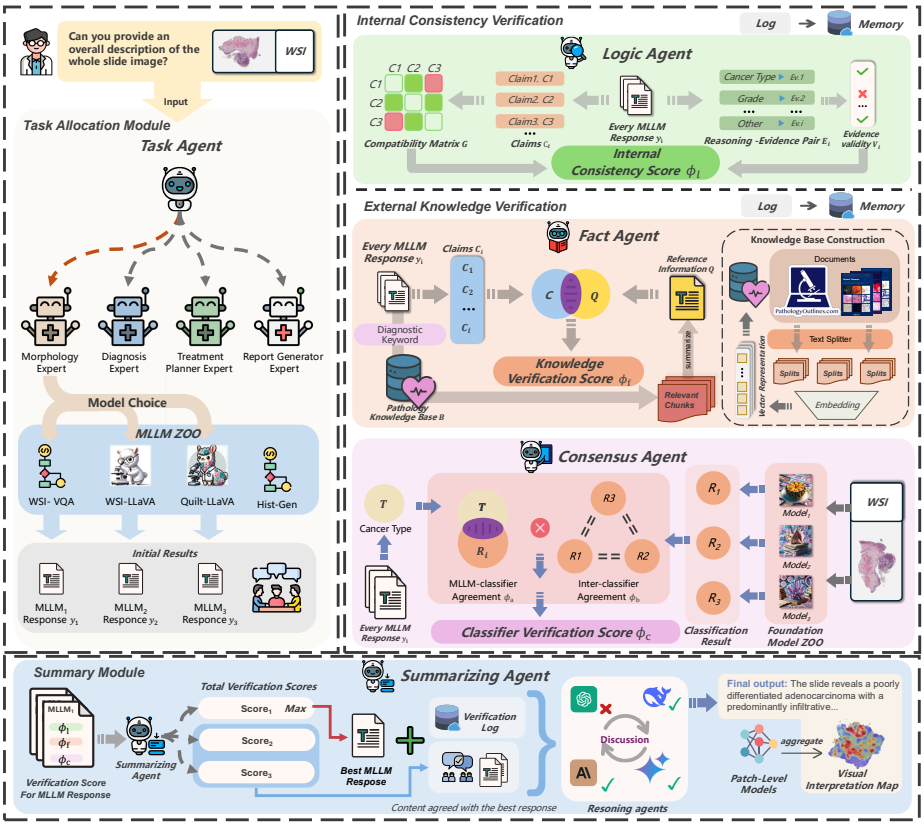



Fig. 2: Overview of **WSI-Agents** for multi-modal WSI analysis with a **task allocation module** for expert allocation, **verification mechanisms** for accuracy assessment, and **summary module** for final summary.

delegates specialized expert agents to generate preliminary responses, effectively addressing various pathology-specific challenges.

**Task Agent.** Specifically, given an input WSI $I$ and question $x$, the task agent first analyzes whether the question relates to WSI analysis, determines the task type of the given question, and identifies which specialized agents $A$ can correctly answer this question based on the task type. The task agent then assigns an appropriate specialized agent to address the question.

**Expert Agent.** We design a series of specialized expert agents to address the primary tasks in WSI analysis, including a morphology expert for analyzing tissue architecture; a diagnosis expert for disease classification; a treatment planning expert for therapeutic recommendations and prognosis assessment; and a report generation expert for generating pathology reports. When assigned a task, the expert agent selects $M$ WSI MLLMs from a pre-defined MLLM model zoo, choosing models that have been trained on relevant tasks and can address the

question, and leverages them to generate initial responses $\{y_i\}_{i=1}^{M}$. We construct the model zoo by integrating five WSI analysis MLLMs [5, 26, 17, 4, 10].

### 2.2 Internal Consistency Verification (ICV)

To ensure the logical consistency of generated responses, we introduce an internal consistency verification system that leverages a logic agent to check for contradictions and validate evidence within responses.

**Logic Agent.** Concretely, the logic agent first extracts the claims $\{C_i\}_{i=1}^{N}$ from the initial responses $y_i$, where $N$ denotes the number of claims. To analyze contradictions within the response, we compute a compatibility matrix $G \in \mathbb{R}^{N\times N}$ representing the compatibility among different claims. Each element $G_{ij} \in [0, 1]$ indicates whether claim $C_i$ can coexist with claim $C_j$. We then compute the compatibility score $\phi_g$ to represent the consistency among claims:

$$\phi_g = \frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N} G_{ij}}{N(N-1)/2}. \tag{1}$$

To further validate each individual claim, the logic agent extracts evidence $\{E_i\}_{i=1}^{N}$ from the response for each claim and assesses the evidence validity $\{V_i\}_{i=1}^{N} \in [0, 1]$ to verify whether evidence $E_i$ supports its corresponding claim $C_i$. The validity score $\phi_e$ is calculated as $\phi_e = \frac{\sum_{i=1}^{N} V_i}{N}$. The final internal consistency score $\phi_l$ is the average of these two scores, $\phi_l = \frac{\phi_g+\phi_e}{2}$.

### 2.3 External Knowledge Verification (EKV)

Due to the limited pathological knowledge in current medical agents [14, 29, 18, 9, 15], we design an external knowledge verification system that employs a fact agent to check for factual errors using a comprehensive knowledge base, alongside a consensus agent that identifies agreement with knowledge embedded in WSI foundation models.

**Fact Agent.** To check for factual errors in generated responses, the fact agent first extracts claims $\{C_i\}_{i=1}^{N}$ from the responses and selects their diagnostic keywords $k$. To obtain relevant knowledge about the input question and response, we establish a comprehensive pathology knowledge base $B$ by collecting resources from pathology websites [24] and authoritative literature [23, 28, 1]. To facilitate efficient retrieval, we segment the documents into splits, extract embeddings for each split, and implement indexing for similarity-based retrieval. Using $k$, we retrieve relevant chunks from $B$ and summarize them into reference information $Q$. To assess factual consistency, we compute a factual score $f_i$ for each claim by determining whether $C_i$ violates the knowledge in $Q$, then average these scores to obtain the knowledge verification score $\phi_k = \frac{\sum_{i=1}^{N} f_i}{N}$.

**Consensus Agent.** Since recent WSI foundation models [8, 20, 21] have acquired extensive knowledge through self-supervised learning on large-scale datasets, we design a consensus agent to leverage this embedded expertise and identify

agreement between these models and generated responses. The consensus agent first extracts the cancer type $T$ from the generated responses. It then draws upon a foundation model zoo comprising several state-of-the-art WSI foundation models (TITAN [8], CONCH [20], and Prism [27]), each demonstrating exceptional classification performance. Next, the agent obtains classification results $\{R_i\}_{i=1}^{H}$ from these models, where $H$ represents the number of classifiers. The consensus agent calculates the agreement $A_i$ between the extracted cancer type $T$ and each classification result $\{R_i\}_{i=1}^{H}$, then computes the MLLM-classifier agreement $\phi_a = \frac{\sum_{i=1}^{H} A_i}{H}$. To account for reliability among different classifiers, the agent also determines the inter-classifier agreement $\phi_b$ using:

$$\phi_b = \frac{2\sum_{i=1}^{H}\sum_{j=i+1}^{H} \mathbb{1}(R_i = R_j)}{H(H-1)}, \tag{2}$$

where $\mathbb{1}(\cdot)$ is an indicator function. The final classifier verification score $\phi_c$ is calculated by using the inter-classifier agreement as a confidence factor: $\phi_c = \phi_a \cdot \phi_b$. The verification logs of logic, fact and consensus agents are stored in the memory module.

### 2.4 Summary Module

**Summarizing Agent.** A summarizing agent synthesizes the final results by unifying all verification scores into a total verification score, $\phi_{total} = w_1\phi_l + w_2\phi_k + w_3\phi_c$, where $w_1$, $w_2$, and $w_3$ represent the importance weights of different verification processes. The agent then identifies the MLLM response with the highest score as the best response and extracts content from other responses that align with this selection. Using this best response, along with the extracted content and verification logs, the summarizing agent generates an initial summary.
**Reasoning Agents.** To further enhance the process, reasoning agents evaluate the initial summary and provide their expert opinions. When disagreements arise, these agents submit specific revision suggestions. The summarizing agent then incorporates this feedback to generate a revised summary for the next round of discussion. This iterative deliberation continues until a consensus is reached—specifically, when more than half of reasoning agents endorse the summary—at which point the discussion concludes and produces final summary. For visual interpretation, we integrate attention maps from multiple patch-level WSI models and synthesize them into a comprehensive visual interpretation map.

## 3 Experiments

**Datasets and Implementation.** We conduct experiments on two main WSI benchmarks, i.e., WSI-Bench [17] and WSI-VQA [5] datasets, and follow their evaluation protocol. We use AutoGen [22] to build our agent framework. For the model with limited input size, all the WSIs are resized to $1024 \times 1024$ thumbnails.For better evaluation of open-ended questions, we employ WSI-Precision,

Table 1: Quantitative comparison on WSI-Bench dataset.

| Methods | Morph. Analysis | | | | | Diagnosis | | | | | Treat. Plan. | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | G.M. | K.D. | R.S. | S.F | Avg. | H.T. | G.R. | M.S. | S.T. | Avg. | T.R. | P.R. | Avg. | |
| Quilt-LLaVA[26] | 0.338 | 0.314 | 0.389 | 0.752 | 0.448 | 0.339 | 0.505 | 0.675 | 0.824 | 0.586 | 0.764 | 0.812 | 0.788 | 0.607 |
| WSI-VQA[5] | 0.322 | 0.313 | 0.389 | 0.554 | 0.395 | 0.377 | 0.430 | 0.388 | 0.550 | 0.436 | 0.708 | 0.874 | 0.791 | 0.541 |
| WSI-LLaVA[17] | 0.390 | 0.350 | 0.450 | 0.760 | 0.488 | 0.410 | 0.570 | 0.630 | 0.830 | 0.610 | 0.790 | 0.830 | 0.810 | 0.325 |
| MDAgents[14] | 0.075 | 0.147 | 0.224 | 0.460 | 0.227 | 0.038 | 0.315 | 0.487 | 0.125 | 0.241 | 0.611 | 0.405 | 0.508 | 0.326 |
| Med-Agents[29] | 0.199 | 0.288 | 0.319 | 0.615 | 0.356 | 0.169 | 0.311 | 0.419 | 0.700 | 0.400 | 0.775 | 0.876 | 0.825 | 0.528 |
| **WSI-Agents** | 0.512 | 0.440 | 0.534 | 0.786 | 0.568 | 0.529 | 0.586 | 0.840 | 0.900 | 0.714 | 0.823 | 0.831 | 0.827 | 0.703 |

Abbreviation: G.M. (Global Morphological Description), K.D. (Key Diagnostic Description), R.S. (Regional Structure Description), S.F. (Specific Feature Description), H.T. (Histological Typing), G.R. (Grading), M.S. (Molecular Subtyping), S.T. (Staging), T.R. (Treatment Recommendations), P.R. (Prognosis).

Table 2: Comparison on report generation task.

| Models | B-1 | B-2 | B-3 | B-4 | R-L | M | Acc |
|---|---|---|---|---|---|---|---|
| Quilt-LLaVA[26] | 0.3343 | 0.1698 | 0.0928 | 0.0566 | 0.2463 | 0.2910 | 0.324 |
| MI-Gen[4] | 0.4027 | 0.3061 | 0.2481 | 0.2085 | 0.4464 | 0.4070 | 0.310 |
| Hist-Gen[10] | 0.4058 | 0.3070 | 0.2482 | 0.2081 | 0.4484 | 0.4162 | 0.300 |
| WSI-LLaVA[17] | 0.3531 | 0.1859 | 0.1058 | 0.0665 | 0.2626 | 0.3072 | 0.380 |
| MDAgents[14] | 0.1287 | 0.0521 | 0.0219 | 0.0118 | 0.1572 | 0.1623 | 0.105 |
| Med-Agents[29] | 0.1610 | 0.0659 | 0.0255 | 0.0130 | 0.1997 | 0.2273 | 0.174 |
| **WSI-Agents** | 0.4443 | 0.3084 | 0.2307 | 0.1825 | 0.4479 | 0.4719 | 0.440 |

*Abbreviations:* B-1/B-2/B-3/B-4 (BLEU-1/2/3/4), R-L (ROUGE-L), M (METEOR), Acc (WSI-Precision [17]).

Table 3: Quantitative Evaluation on WSI-VQA.

| Models | Acc |
|---|---|
| Quilt-LLaVA[26] | 0.130 |
| WSI-VQA[5] | 0.470 |
| WSI-LLaVA[17] | 0.550 |
| MDAgents[14] | 0.208 |
| Med-Agents[29] | 0.183 |
| **WSI-Agents** | 0.600 |

*Abbreviations:* Acc (Accuracy).

which compares extracted claims from ground-truth with model outputs and averages the scores; this metric is used in Table 1 and as the accuracy (Acc) measure in Table 2.

**Results on WSI-Bench.** We evaluate WSI-Agents against WSI MLLMs [26, 5, 17] and medical agents [14, 29] on the WSI-Bench dataset [17]. Table 1 demonstrates that WSI-Agents achieves superior performance across all three main tasks, surpassing Quilt-LLaVA [26] and Med-Agents [29] by substantial margins of approximately 10% and 17%, respectively. For the report generation task, Table 2 shows that WSI-Agents outperforms WSI-LLAVA [17] by 6% in accuracy. These results clearly indicate that WSI-Agents surpasses existing WSI MLLMs and medical agents in performance. For qualitative comparison, Table 4 shows other methods misclassify the sample despite identifying some correct features, while WSI-Agents correctly identifies adenocarcinoma with accurate supporting evidence, demonstrating superior diagnostic precision in WSI interpretation.

**Results on WSI-VQA Dataset.** In Table 3, our method achieves 60.0% accuracy, outperforming previous approaches including WSI-LLaVA [17] (55.0%), WSI-VQA [5] (47.0%), and Quilt-LLaVA [26] (13.0%). Medical agents like MDAgents [14] (20.8%) and Med-Agents [15] (18.3%) performed notably worse, suggesting general medical knowledge is insufficient for WSI analysis. These results highlight the superiority of WSI-Agents through its effective integration of pathology knowledge and verification mechanism.

Table 4: Visual comparison on WSI-Bench. Green for correct and Red for wrong.

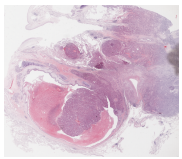
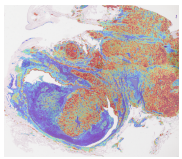
| | |
|---|---|
| **Image** |  Input WSI  Generated Visual Interpretation Map |
| **Question** | What is the histological classification based on your examination of the slide? |
| **Quilt-LLaVA** | The histological classification is seminoma. This is supported by the presence of large, uniform cells with clear cytoplasm and prominent nucleoli, arranged in nests and sheets, and the characteristic fibrous stroma with lymphocytic infiltration. |
| **WSI-VQA** | The histological classification based on the observed features is invasive squamous cell carcinoma . This classification is supported by the presence of poorly differentiated squamous cells with marked nuclear pleomorphism and high mitotic activity. |
| **WSI-LLaVA** | The histological classification is high-grade urothelial carcinoma. This classification is supported by the presence of poorly differentiated tumor cells, significant nuclear atypia, high mitotic activity, and evidence of lymphovascular invasion. |
| **WSI-Agents** | The histological classification...is adenocarcinoma. This classification is consistent with the identification of glandular structures,...poorly differentiated cells with nuclear pleomorphism and high mitotic activity. Additionally,...keratinization or intercellular bridges...are absent in the provided image.There is no evidence of lymphovascular invasion or specific stromal characteristics...were noted in the image, their absence aligns with the adenocarcinoma classification rather than alternatives like squamous cell carcinoma, high-grade urothelial carcinoma or seminoma. |
| **Ground Truth** | Based on the observed features, the histological classification is adenocarcinoma. The poorly differentiated nature of the glandular epithelial cells, along with marked pleomorphism and significant mitotic activity, supports this classification. The invasive pattern further confirms the diagnosis of adenocarcinoma. |

Table 5: Ablation studies on WSI-Bench dataset.

| TAM | | ICV | EKV | | Summary | | Morph. | Diagnosis | Treat. | Report | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Task | Expert | Logic | Fact | Cons. | Summ. | Reas. | Analysis | | Plan. | Gen. | |
| ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.362 | 0.560 | 0.786 | 0.415 | 0.531 |
| ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.289 | 0.427 | 0.528 | 0.321 | 0.391 |
| ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ | 0.539 | 0.648 | 0.817 | 0.424 | 0.607 |
| ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | 0.521 | 0.647 | 0.814 | 0.431 | 0.603 |
| ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | ✓ | 0.531 | 0.644 | 0.814 | 0.418 | 0.601 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ | 0.416 | 0.586 | 0.810 | 0.332 | 0.536 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | 0.469 | 0.607 | 0.806 | 0.362 | 0.561 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.568 | 0.714 | 0.827 | 0.440 | 0.637 |

Abbreviations: TAM (Task Allocation Module), ICV (Internal Consistency Verification), EKV (External Knowledge Verification), Cons. (Consensus), Summ. (Summarizing), Reas. (Reasoning), Morph. (Morphological), Treat. Plan. (Treatment Planning), Report Gen. (Report Generation), Avg. (Average).

**Ablation Studies.** In Table 5, we conduct ablation studies on the WSI-Bench dataset to evaluate the effectiveness of each module, including the task allocation module (TAM), internal consistency verification (ICV), external knowledge verification (EKV) and summary module (summary). When ablating each agent in the corresponding module, the performance across all the tasks decreases significantly, indicating the effectiveness of each agent.

## 4 Conclusion

We present a collaborative multi-agent framework (WSI-Agents) for multi-modal WSI analysis by integrating specialized agents and verification mechanisms. Through task allocation, consistency checks, knowledge verification, and summarization, WSI-Agents improves accuracy, adaptability, and interpretability for pathological tasks. Experiments demonstrate superior performance over existing WSI MLLMs and current medical agents. Ablation studies prove the effectiveness of each module.

**Acknowledgments.** This study was supported by the National Natural Science Foundation of China (Grants 82261138629 and 12326610), the Guangdong Provincial Key Laboratory (Grant 2023B1212060076), the Yongjiang Technology Innovation Project (Grant 2022A-097-G), the Zhejiang Department of Transportation General Research and Development Project (Grant 2024039), and the National Natural Science Foundation of China talent grant (UNNC: B0166).

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

## References

1. Bosman, F.T., Carneiro, F., Hruban, R.H., Theise, N.D.: WHO classification of tumours of the digestive system. No. Ed. 4 (2010)
2. Bulten, W., Pinckaers, H., Van Boven, H., Vink, R., De Bel, T., Van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G.: Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology **21**(2), 233–241 (2020)
3. Çayır, S., Darbaz, B., Solmaz, G., Yazıcı, Ç., Kusetogulları, H., Tokat, F., Iheme, L.O., Bozaba, E., Tekin, E., Özsoy, G., et al.: Patch-based approaches to whole slide histologic grading of breast cancer using convolutional neural networks. In: Diagnostic Biomedical Signal and Image Processing Applications with Deep Learning Methods, pp. 103–118. Elsevier (2023)
4. Chen, P., Li, H., Zhu, C., Zheng, S., Shui, Z., Yang, L.: Wsicaption: Multiple instance generation of pathology reports for gigapixel whole-slide images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 546–556. Springer (2024)
5. Chen, P., Zhu, C., Zheng, S., Li, H., Yang, L.: Wsi-vqa: Interpreting whole slide images by generative visual question answering. In: European Conference on Computer Vision. pp. 401–417. Springer (2024)
6. Chen, Y., Wang, G., Ji, Y., Li, Y., Ye, J., Li, T., Zhang, B., Pei, N., Yu, R., Qiao, Y., et al.: Slidechat: A large vision-language assistant for whole-slide pathology image understanding. arXiv preprint arXiv:2410.11761 (2024)
7. Ciga, O., Xu, T., Nofech-Mozes, S., Noy, S., Lu, F.I., Martel, A.L.: Overcoming the limitations of patch-based learning to detect cancer in whole slide images. Scientific Reports **11**(1), 8894 (2021)
8. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., et al.: Multimodal whole slide foundation model for pathology. arXiv preprint arXiv:2411.19666 (2024)

9. Fallahpour, A., Ma, J., Munim, A., Lyu, H., Wang, B.: Medrax: Medical reasoning agent for chest x-ray. arXiv preprint arXiv:2502.02673 (2025)
10. Guo, Z., Ma, J., Xu, Y., Wang, Y., Wang, L., Chen, H.: 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. pp. 189–199. Springer (2024)
11. Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H., Takeuchi, I.: Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In: CVPR. pp. 3852–3861 (2020)
12. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2424–2433 (2016)
13. Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G., Srinivasan, B.: A generalized deep learning framework for whole-slide image segmentation and analysis. Scientific reports **11**(1), 11579 (2021)
14. Kim, Y., Park, C., Jeong, H., Chan, Y.S., Xu, X., McDuff, D., Lee, H., Ghassemi, M., Breazeal, C., Park, H.W.: Mdagents: An adaptive collaboration of llms in medical decision making. arXiv preprint arXiv:2404.15155 (2024)
15. Li, B., Yan, T., Pan, Y., Luo, J., Ji, R., Ding, J., Xu, Z., Liu, S., Dong, H., Lin, Z., et al.: Mmedagent: Learning to use medical tools with multi-modal agent. arXiv preprint arXiv:2407.02483 (2024)
16. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., Tu, Z.: Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118 (2023)
17. Liang, Y., Lyu, X., Ding, M., Chen, W., Zhang, J., Ren, Y., He, X., Wu, S., Yang, S., Wang, X., et al.: Wsi-llava: A multimodal large language model for whole slide image. arXiv preprint arXiv:2412.02141 (2024)
18. Liu, J., Wang, W., Ma, Z., Huang, G., SU, Y., Chang, K.J., Chen, W., et al.: Medchain: Bridging the gap between llm agents and clinical practice through interactive sequential benchmarking. arXiv preprint arXiv:2412.01605 (2024)
19. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., et al.: Agentbench: Evaluating llms as agents. arXiv preprint arXiv:2308.03688
20. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**(3), 863–874 (2024)
21. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: CVPR. pp. 19764–19775 (2023)
22. Microsoft: Autogen. https://github.com/microsoft/autogen (2023), 2025-02-27
23. Nagtegaal, I.D., Odze, R.D., Klimstra, D., Paradis, V., Rugge, M., Schirmacher, P., Washington, K.M., Carneiro, F., Cree, I.A., et al.: The 2019 who classification of tumours of the digestive system. Histopathology **76**(2), 182 (2019)
24. PathologyOutlines: PathologyOutlines.com (2001), 2025-02-27
25. Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., Sun, M.: Communicative agents for software development. arXiv preprint arXiv:2307.07924 (2023)
26. 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 CVPR (2024)

27. Shaikovski, G., Casson, A., Severson, K., Zimmermann, E., Wang, Y.K., et al.: Prism: A multi-modal generative foundation model for slide-level histopathology. arXiv preprint arXiv:2405.10254 (2024)
28. Tan, P.H., Ellis, I., Allison, K., Brogi, E., Fox, S.B., Lakhani, S., Lazar, A.J., Morris, E.A., Sahin, A., Salgado, R., et al.: The 2019 who classification of tumours of the breast. Histopathology **77**(2) (2020)
29. Tang, X., Zou, A., Zhang, Z., Li, Z., Zhao, Y., Zhang, X., Cohan, A., Gerstein, M.: Medagents: Large language models as collaborators for zero-shot medical reasoning. arXiv preprint arXiv:2311.10537 (2023)
30. Tripathi, S., Singh, S.K., Lee, H.K.: An end-to-end breast tumour classification model using context-based patch modelling–a bilstm approach for image classification. Computerized Medical Imaging and Graphics **87**, 101838 (2021)
31. Wang, H., Zhao, S., Qiang, Z., Xi, N., Qin, B., Liu, T.: Beyond direct diagnosis: Llm-based multi-specialist agent consultation for automatic diagnosis. arXiv preprint arXiv:2401.16107 (2024)
32. 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 pp. 1–8 (2024)