# Clinically-inspired spatial reasoning on gigapixel cancer images via think-act-reflect loops

Anonymous CVPR submission

Paper ID *****

## Abstract

*Deciphering tumor microenvironment from whole slide images (WSIs) is intriguing as it holds the key to cancer diagnosis, prognosis and treatment response. While these gigapixel images on one hand offer a comprehensive portrait of cancer, on the other hand, the extremely large size, as much as more than 10 billion pixels, make it challenging and time-consuming to navigate to corresponding regions to support diverse clinical inspection. Inspired by pathologists who conducted navigation on WSIs with a combination of sampling, reasoning and self-reflection, we proposed "GigaVisualReasoning", a multi-modal reasoning agent based on GPT-4o that iteratively navigates through multiple rounds of reasoning and refinements. Concretely, starting with randomly sampled candidate regions, GigaVisualReasoning reviews current selections with self-reflection, reasoning over the correspondence between visual observations and clinical questions, and conclude by proposing new regions to explore. Across rounds, GigaVisualReasoning builds a reasoning chain that gradually narrows attention to diagnostically relevant areas. We evaluated our method on cancer subtyping and visual question answering (VQA) tasks, comparing against majority vote selection and GPT baselines. Our results show that iterative multi-modal reasoning achieved at 8.5% improvements on classification accuracy on average. Moreover, the agentic system provides a transparent decision trajectory, making the exploration process interpretable and can provide more insights into clinical decision making.*

## 1. Introduction

The Whole Slide Images (WSIs) for pathology serves as the gold standard for cancer diagnosis. Deciphering the tumor microenvironment from WSIs is of paramount importance, which informs the current cancer stage, potential treatment therapy and future treatment response. These gigapixel images are often extremely large, as much as more than 10 billion pixels, making it difficult for navigating to the corresponding regions that need to be checked for clinical decision making [8]. In practice, they are so large that even experienced pathologists could not analyze the entire slide at once. Instead, pathologists navigate through the whole slide image by sampling and carefully examining several specific regions of interest (ROIs). If the current regions could not well reflect the cancer development for making clinical decisions, they will propose a new set of regions which are more likely to contain prominent signals based on their experience [6]. In summary, the critical component of this process is building a long reasoning process to navigate on WSIs.

However, most existing computational methods for analyzing WSIs [1, 4, 9] do not have the built-in reasoning capability for navigation. They either rely on fixed-size patches extracted uniformly or on heuristic-based region selection. These approaches often include irrelevant areas and may miss important features [15]. In addition, many current pipelines process selected ROIs in a single step without considering how pathologists refine their focus across multiple rounds through thinking and reflection [17]. As a result, these approaches suffer from selecting the most informative ROI, which is often the most critical step in downstream tasks such as subtyping or question answering [5].

Recent developments in multi-modal large language models (MLLMs) have enabled integrated visual and language reasoning [3], making them suitable for tasks such as interpreting pathology images through natural language prompts [12, 13, 16]. These models can be guided through reasoning steps, making them suitable for simulating decision-making processes [14]. Rather than choosing a region in a single step, MLLMs can be prompted to revise and justify their selection iteratively.

Therefore, we propose GigaVisualReasoning, as shown in Figure 1, which uses a multi-modal large language model (GPT-4o) to simulate the reasoning process that a human expert uses for navigation on gigapixel images [10]. We design special prompts to expose the multi-modal reasoning capability of frontier models for navigation. Our model

employs an iterative think-act-reflect loop to mimic the internal reasoning process of human experts. Specifically, for each clinical question, each loop consists of three steps: (1) GigaVisualReasoning adaptively infers ROIs by associating the visual observations to clinical knowledge, (2) GigaVisualReasoning automatically retrieves corresponding regions by specifying the coordinates and magnification level of regions, (3) GigaVisualReasoning refine the chosen region with self-reflection on whether it contains sufficient information for clinical question answering. With the built-in reasoning capability in GigaVisualReasoning, we observed significant improvements compared to competing approaches, including 8.5% improvements on 10 cancer subtyping tasks and 11.16% improvements on VQA tasks across 7 cancer types. Our experiments demonstrated improved performance as more reasoning steps are deployed, echoing the recent progress of LLMs on test-time scaling [11].

# 2. Methods

## 2.1. GigaVisualReasoning with Iterative Selection

We develop GigaVisualReasoning to emulate the reasoning process of a pathologist navigating a whole slide image (WSI). Rather than selecting a region in a single step, the system performs multiple rounds of selection, reflection, and refinement. In each round, GPT-4o receives a set of candidate regions of interest (ROIs) and is prompted to select the one most informative for the clinical task. Along with the selection, GPT-4o is also asked to justify its choice based on visual and contextual features.

After the third round, the system transitions to a free-form reasoning mode, where GPT-4o proposes new coordinates directly, without being constrained to fixed candidates. This forms an iterative *think–act–reflect loop*: the model reasons over the visual and textual context (*think*), selects a new region (*act*), and evaluates whether the new choice improves task relevance (*reflect*). All previously selected ROIs and justifications are retained and passed into subsequent rounds, allowing the model to accumulate and reason over a growing body of context, simulating how a human expert refines attention across rounds.

The loop terminates when GPT-4o determines that further refinement will not improve task performance, either based on a fixed round limit or an explicit stop condition. The final selected ROI is then used to perform the downstream task, such as cancer subtyping or report generation, using a task-specific prompt. This approach supports transparent decision trajectories and adapts flexibly to multiple clinical endpoints. Compared to one-step or static ROI selection methods, GigaVisualReasoning allows the model to correct earlier choices and iteratively converge on more diagnostic regions, demonstrating improved alignment with expert behavior.

## 2.2. Baselines

We compare our method against two single-step baselines that bypass iterative reasoning. In the *Majority-Vote Baseline*, the system randomly selects 21 non-empty regions from the WSI. Each region is passed to GPT-4o for prediction, and the final output is determined by majority vote among the 21 predictions. This approach serves as a reference for random selection without any refinement.

In the *GPT Baseline*, the system samples 20 random non-empty regions and presents all of them to GPT-4o at once. The model selects the region it finds most relevant to the task, and that single region is used for prediction.

# 3. Experiments

## 3.1. Dataset

We evaluate our method on WSIs from The Cancer Genome Atlas (TCGA) project. All slides are hematoxylin and eosin (H&E)-stained and scanned at ×40 magnification.

For the cancer subtyping task, we use WSIs from 10 organs: Breast (BRCA), Bowel (COLON), Esophagus/Stomach (ESO), Biliary Tract/Liver (HEP), Lung (LUNG), Kidney (RCC), CNS/Brain (GLIOMA), Adrenal gland (ADREN), Cervix (CERVIX), and Pleura (PLEURA). Each slide is paired with a ground truth subtype label selected from a predefined list of clinically meaningful cancer subtypes [15].

For the visual question answering (VQA) task, we use diagnostic reports sourced from the TCGA Pathology Reports dataset (https://github.com/tatonetti-lab/tcga-path-reports). These reports provide detailed clinical observations for WSIs and serve as reference texts for evaluating the quality and relevance of generated responses [7].

## 3.2. Tasks

We evaluate our method on two tasks: cancer subtyping and visual question answering (VQA). Both tasks are performed on selected regions of interest (ROIs) extracted from whole slide images (WSIs). The selection process differs across methods, as described earlier. For each task, we measure performance using appropriate metrics that reflect either classification accuracy or the quality of generated text.

### 3.2.1. Cancer Subtyping

In the cancer subtyping task, the model is asked to classify each WSI into one of several predefined cancer subtypes. The input to the model is a single ROI selected from the slide, and the output is a predicted subtype label. We use classification accuracy for each cancer type as the evaluation metric, comparing the model's predicted label with the

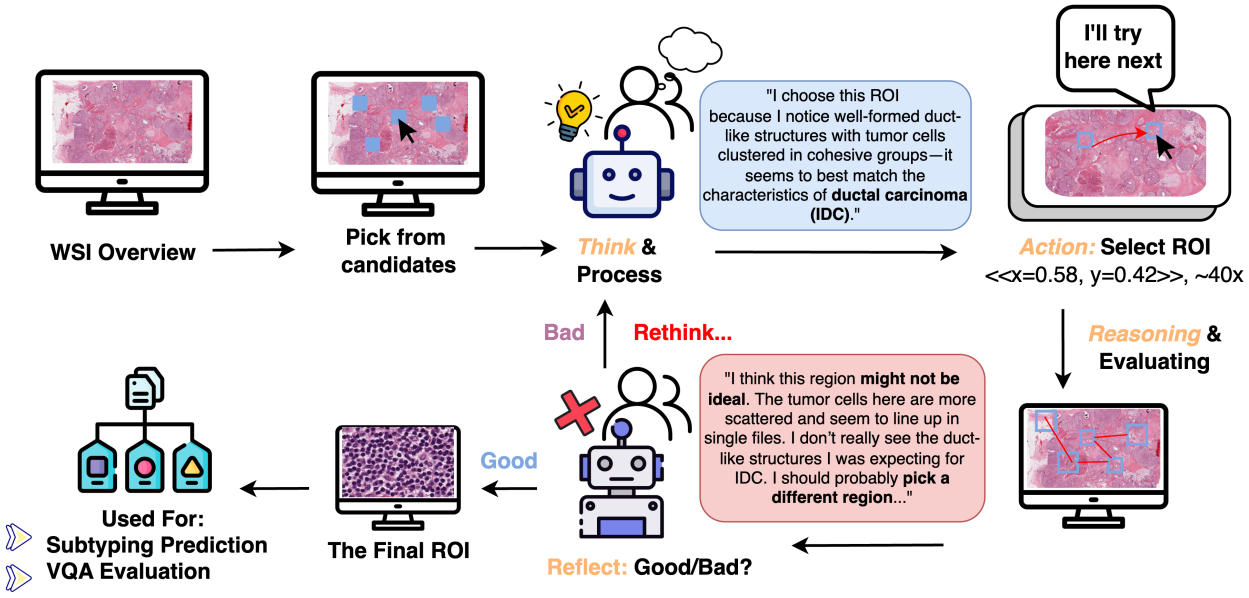
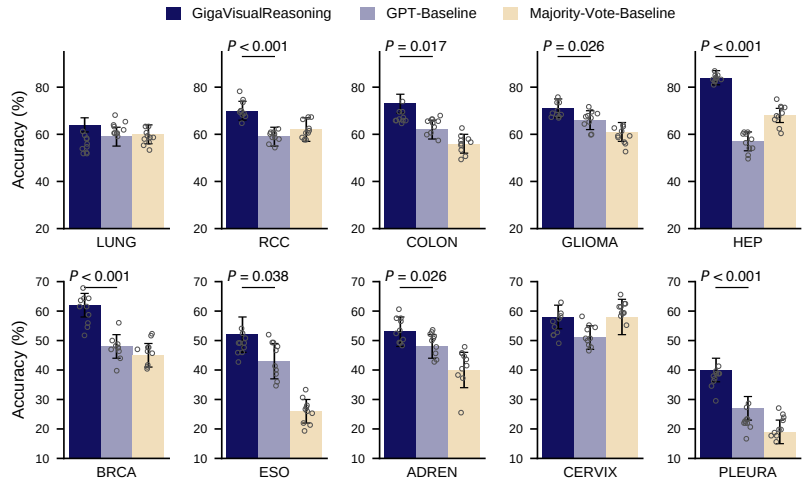
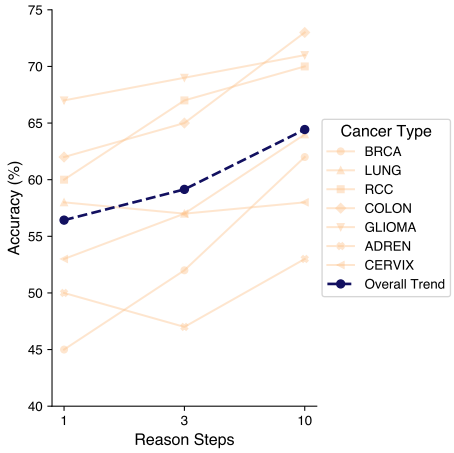



Figure 1. An illustration of our ROI-Agent reasoning process. The agent starts from WSI overview, proposes candidates, and iteratively selects ROIs through a loop of thinking, acting, and reflecting. The final selected ROI is used for downstream tasks such as cancer subtyping or VQA evaluation.




(a) Accuracy comparison across 10 cancer types for subtyping. GigaVisualReasoning consistently outperforms both baselines with statistically significant improvements ($p < 0.05$).




(b) Impact of reasoning iterations on accuracy. Performance improves as the agent refines ROI selection, plateauing after 3 rounds.

Figure 2. Quantitative results on the cancer subtyping task. (a) compares classification accuracy across cancer types, and (b) illustrates how multi-round reasoning improves performance.

known ground truth subtype. No additional clinical metadata is provided during inference.

#### 3.2.2. Visual Question Answering (VQA)

For the VQA task, we assess whether the selected ROI can support meaningful medical interpretation. After the ROI is chosen, we prompt GPT-4o to generate a short scientific report describing the ROI. The prompt includes a few in-context examples, taken from real diagnostic reports for similar slides. The generated report is compared to the reference report using a GPT-based evaluation score (from 0 to 10, where 10 stands for the perfect match), which measures token overlap and semantic similarity [2]. This score serves as a proxy for clinical relevance and completeness of the generated text.

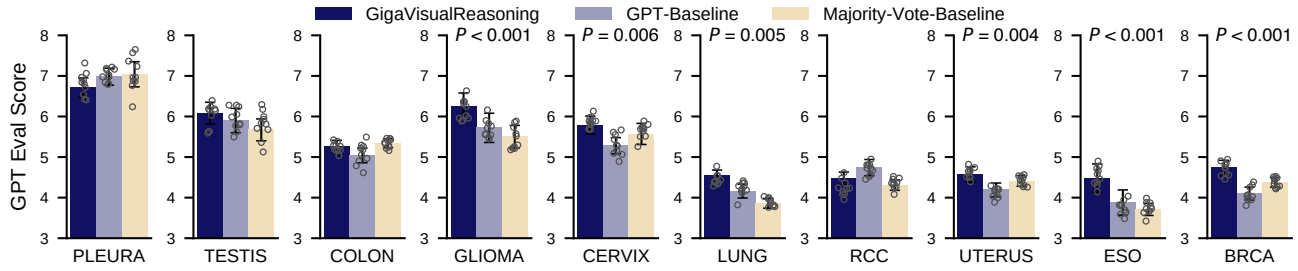



Figure 3. GPT-based evaluation scores for the generated diagnostic reports across 10 cancer types. GigaVisualReasoning achieves higher GPT Eval scores on most categories, indicating improved clinical relevance and completeness of the generated responses.

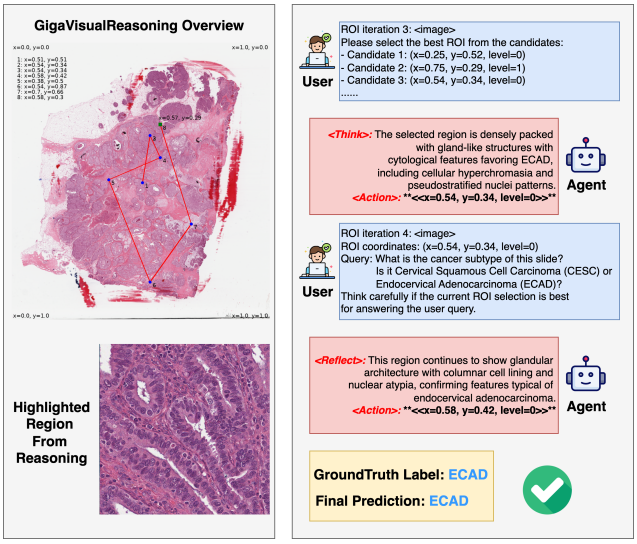



Figure 4. Case study on the ECAD subtype. The ROI-Agent iteratively selects regions by reviewing visual candidates and generating feedback.

### 3.3. Evaluation Metrics

We evaluate our method using task-specific metrics for cancer subtyping and VQA. For the cancer subtyping task, we report the subtyping accuracy for each cancer type. For the VQA task, we assess the quality of the generated diagnostic reports using a GPT-based evaluation score. This score ranges from 0 to 10, where higher values indicate better token-level and semantic-level similarity between the generated report and the ground truth clinical report.

## 4. Results

### 4.1. Subtyping Performance

We evaluate cancer subtype prediction across 10 cancer types by comparing our ROI-Agent to the GPT-Baseline and the Majority-Vote Baseline. Figure 2a shows the accuracy for each cancer type. Our ROI-Agent consistently achieves higher accuracy than the baselines across all datasets, with statistically significant improvements ($p < 0.05$) for most cancer types.

### 4.2. Iterative Reasoning Performance

We further investigate the impact of multi-round reasoning by evaluating model performance across different numbers of iterations. Figure 2b illustrates the trend of accuracy as the number of reasoning rounds increases. We observe that accuracy improves steadily as the agent performs more reasoning steps, reaching a plateau after approximately 3 rounds.

### 4.3. VQA Performance

For the VQA task, we assess the quality of generated diagnostic reports using a GPT-based evaluation score ranging from 0 to 10. Figure 3 presents the VQA performance across cancer types. Our ROI-Agent achieves higher GPT Eval Scores compared to the GPT-Baseline and the Majority-Vote Baseline for most cancer types.

### 4.4. Case Study: ECAD/CESC Subtyping

Figure 4 shows an example from the CESC dataset, where the ROI-Agent is asked to predict the cancer subtype. The correct label is Endocervical Adenocarcinoma (ECAD).

At iteration 3, the agent selects the region at (x = 0.54, y = 0.34), noting gland-like structures with hyperchromasia, a feature typical of ECAD. In the next round, based on updated context, it shifts to (x = 0.58, y = 0.42), where it observes clearer glandular organization and nuclear atypia with columnar lining—more consistent with the expected morphology.

Using this refined region, the agent correctly predicts the ground-truth label. This example illustrates how the model iteratively improves its decision by revisiting earlier steps. Each round focuses on task-relevant histologic features, and the reasoning path remains interpretable for verifying the prediction.

# References

[1] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 1

[2] 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. 3

[3] Bin Feng, Zequn Liu, Nanlan Huang, Zhiping Xiao, Haomiao Zhang, Srbuhi Mirzoyan, Hanwen Xu, Jiaran Hao, Yinghui Xu, Ming Zhang, et al. A foundation model for bioactivity prediction using pairwise meta-learning. *bioRxiv*, pages 2023–10, 2023. 1

[4] Fatemeh Ghezloo, Oliver H Chang, Stevan R Knezevich, Kristin C Shaw, Kia Gianni Thigpen, Lisa M Reisch, Linda G Shapiro, and Joann G Elmore. Robust roi detection in whole slide images guided by pathologists' viewing patterns. *Journal of Imaging Informatics in Medicine*, 38(1): 439–454, 2025. 1

[5] Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom O Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G Elmore, Ranjay Krishna, and Linda Shapiro. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. *arXiv preprint arXiv:2502.08916*, 2025. 1

[6] Ekta Jain, Ankush Patel, Anil V Parwani, Saba Shafi, Zoya Brar, Shivani Sharma, and Sambit K Mohanty. Whole slide imaging technology and its applications: Current and emerging perspectives. *International Journal of Surgical Pathology*, 32(3):433–448, 2024. 1

[7] Jenna Kefeli and Nicholas Tatonetti. Tcga-reports: A machine-readable pathology report resource for benchmarking text-based ai models. *Patterns*, 5(3), 2024. Published March 8, 2024. 2

[8] Kechun Liu. *Interpretable Analysis of Melanoma in Whole Slide Imaging: Detection, Virtual Staining, and Diagnostic Insights*. PhD thesis, University of Washington, 2025. 1

[9] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 1

[10] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033): 466–473, 2024. 1

[11] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025. 2

[12] Eleni Skourti. A vision–language foundation model for clinical oncology. *Nature Cancer*, pages 1–1, 2025. 1

[13] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, 30 (10):2924–2935, 2024. 1

[14] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035): 970–978, 2024. 1

[15] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. 1, 2

[16] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025. 1

[17] Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, 22(1):166–176, 2025. 1