MEDSPACEAGENT: TRAINING-FREE TOOL-GROUNDED SPATIAL REASONING FOR MEDICAL VISUAL QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical visual question answering (VQA) models often produce answers through implicit cross-modal reasoning, making it difficult to determine whether their predictions are supported by relevant image regions and valid anatomical relationships. In this paper, we introduce **MedSpaceAgent**, a training-free, tool-grounded agent for explicit spatial reasoning in Medical VQA. Instead of directly generating an answer, MedSpaceAgent translates each question into an executable spatial program and dynamically invokes frozen medical perception tools for anatomical segmentation, abnormality localization, and region-level inspection. The resulting observations are organized into a spatial evidence graph, on which a deterministic executor computes anatomical, directional, and distance-based relations. An evidence-aware verifier further assesses evidence completeness, geometric consistency, and agreement across tools, enabling the agent to replan its perception strategy. This design separates linguistic planning from spatial computation and provides a reasoning trajectory without updating any foundation-model parameters.The proposed framework establishes a transparent alternative to end-to-end adaptation and provides a general foundation for reliable spatial reasoning in medical multimodal systems.

1 INTRODUCTION

Medical visual question answering (Medical VQA) aims to answer natural-language questions about medical images and provides an intuitive interface for clinical image interpretation, medical education, and decision support Lau et al. (2018); He et al. (2020); Liu et al. (2021); Lin et al. (2023). Recent medical vision-language models have demonstrated promising performance by leveraging large-scale vision–language pretraining and instruction tuning Huang et al. (2021); Chen et al. (2022); Wu et al. (2023a); Li et al. (2023a). However, reliable Medical VQA requires more than recognizing abnormalities or associating visual patterns with diagnostic terms. Many clinically relevant questions rely on the spatial organization of anatomical structures and pathological findings, such as whether a lesion is located in the left or right organ, whether it lies above or below an anatomical landmark, whether it overlaps an adjacent structure, or which abnormality is closer to a clinically important boundary Tripathi et al. (2026); Monon et al. (2026). Answering such questions requires the system to identify the relevant entities, localize them accurately, and explicitly reason about their anatomical and geometric relationships.

Existing Medical VQA methods Nguyen et al. (2019); Zhan et al. (2020); Khare et al. (2021); Li et al. (2023b) predominantly formulate the task as an end-to-end mapping from an image–question pair to an answer, i.e., $A = f_\theta(I, Q)$, where both visual perception and spatial reasoning are implicitly encoded in the model parameters. Although task-specific fine-tuning with VQA examples or region annotations can improve benchmark performance, it does not guarantee that an answer is supported by the relevant image regions or derived from a valid spatial relationship. A model may exploit linguistic priors and dataset-specific correlations, produce the correct answer through an unfaithful reasoning process, or remain overconfident when the required visual evidence is unavailable. Moreover, spatial relationships embedded in latent representations are difficult to inspect and independently verify. Recent medical agents provide an alternative paradigm by allowing foundation models to interact with external segmentation, localization, retrieval, and measurement tools Li

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

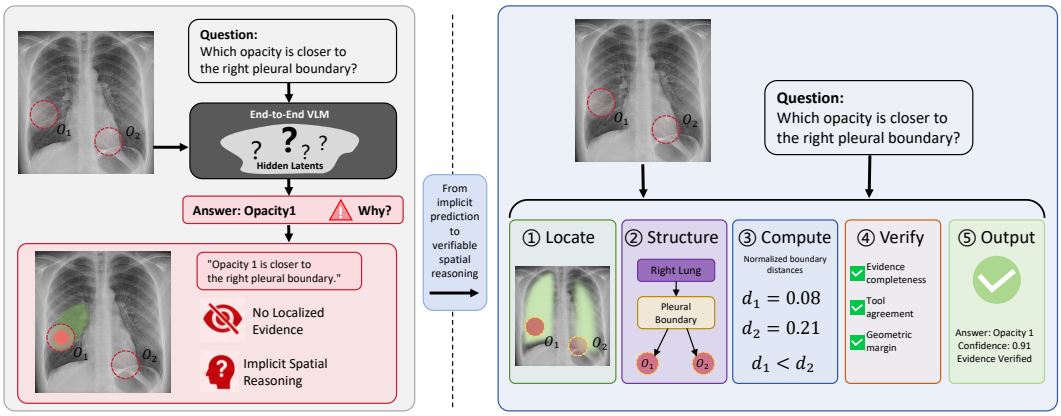

Figure 1: **Motivation for verifiable spatial reasoning in Medical VQA.** Conventional end-to-end models may produce a plausible answer from implicit representations or diffuse visual attention, but they do not establish whether the answer follows from the relevant anatomical regions and spatial relationships. In contrast, MedSpaceAgent explicitly localizes the target entities, structures the acquired evidence, computes the requested geometric relation, and verifies evidence completeness and consistency before answering.

et al. (2024); Fallahpour et al. (2025); Guo et al. (2025). Nevertheless, most existing systems either invoke tools through fixed pipelines or convert tool outputs into textual context, without ensuring that the final answer is geometrically consistent with the acquired visual evidence.

To address these limitations, we propose **MedSpaceAgent**, a training-free, tool-grounded spatial reasoning agent for Medical VQA. Figure 1 illustrates the central motivation of our work. Instead of directly generating an answer, MedSpaceAgent translates the question into an executable spatial program that specifies the target entities, anatomical constraints, required evidence, and spatial operations. The agent then dynamically invokes frozen medical perception tools to segment anatomical structures, localize abnormalities, inspect relevant regions, and verify image orientation. The resulting observations are organized into a spatial evidence graph that records entities, anatomical coordinates, quantitative relations, confidence values, and tool provenance. A deterministic spatial executor operates on this graph to compute directional, topological, distance-based, and comparative relationships. This design separates linguistic planning from spatial computation: the language model decides which operations are required, whereas spatial facts are established exclusively through validated tool outputs and explicit geometric calculations.

In summary, our contributions are as follows:

- We propose **MedSpaceAgent**, a training-free, tool-grounded framework that reformulates Medical VQA as an explicit process of evidence acquisition, spatial reasoning, and verification.
- We introduce a question-conditioned orchestration mechanism that translates medical questions into executable spatial programs and dynamically selects relevant frozen perception tools.
- We construct a spatial evidence graph and a deterministic executor to compute spatial relations from localized entities, making each answer traceable to visual evidence and quantitative measurements.

2 Related Work

2.1 Medical Visual Question Answering

Medical visual question answering (Medical VQA) requires models to jointly understand medical images, clinical questions, and domain knowledge. Early studies introduced task-specific datasets, including VQA-RAD Lau et al. (2018), PathVQA He et al. (2020), SLAKE Liu et al. (2021), and

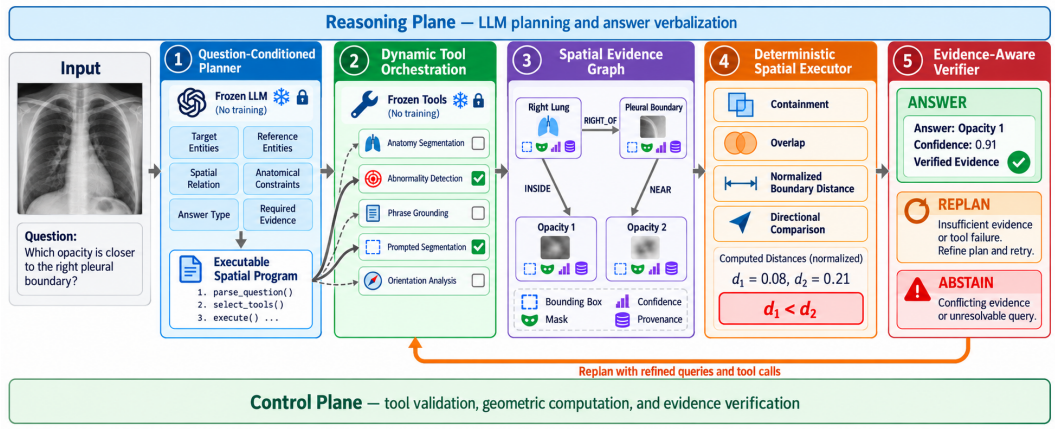

Figure 2: **Overview of MedSpaceAgent.** Given a medical image and a question, a frozen language model parses the spatial intent and generates an executable program that dynamically selects the required frozen medical perception tools. Their validated outputs are organized into a spatial evidence graph, on which a deterministic executor computes containment, overlap, directional, and normalized distance relations. An evidence-aware verifier then evaluates evidence completeness, geometric consistency, and cross-tool agreement to determine whether the system should answer, replan with refined tool calls, or abstain. The reasoning plane performs linguistic planning and answer verbalization, whereas the control plane validates tool observations, executes geometric operations, and enforces evidence-grounded decisions.

PMC-VQA Zhang et al. (2023). Subsequent methods employed conditional reasoning, data-efficient learning, and external medical knowledge Zhan et al. (2020); Nguyen et al. (2019); Do et al. (2021); Gong et al. (2022); Wu et al. (2023a). Medical vision–language pre-training further improved cross-modal representations through global–local alignment, masked modeling, and contrastive objectives Huang et al. (2021); Boecking et al. (2022); Chen et al. (2022); Li et al. (2023c;b). More recently, instruction-tuned multimodal models have supported open-ended biomedical question answering Li et al. (2023a); Chen et al. (2024); Deria et al. (2026); Bai et al. (2025); Xu et al. (2026). Nevertheless, most existing methods leave visual perception and reasoning implicit in model parameters; even rationale-guided prediction Gai et al. (2025) does not by itself guarantee that an answer follows from localized evidence. In contrast, our method performs no task-specific fine-tuning and solves Medical VQA through active evidence acquisition, explicit spatial computation, and answer verification.

2.2 Spatial Reasoning in Medical Images

Spatial reasoning is essential for interpreting the anatomical locations of abnormalities and their relationships to surrounding structures. Existing medical vision-language approaches have partially addressed this capability through global–local representation learning Huang et al. (2021), region–phrase alignment Chen et al. (2023), medical grounding Deria et al. (2026); Yu et al. (2026), and topology-aware representation learning Meng et al. (2026). Recent benchmarks expose persistent weaknesses in two-dimensional, three-dimensional, and radiology-specific spatial understanding Jia et al. (2025); Monon et al. (2026); Tripathi et al. (2026). However, localizing an individual entity does not necessarily establish its relationship to another entity. Determining whether a lesion lies inside an organ, overlaps an anatomical structure, or is closer to one landmark than another requires explicit multi-entity reasoning and geometric computation. Training-free visual programming and language-to-space methods demonstrate the value of decomposing visual reasoning into executable operations Gupta & Kembhavi (2023); Surís et al. (2023); Mi et al. (2025), while recent spatial agents emphasize explicit actions and geometric constraints Chen et al. (2025b); Cho et al. (2026); Jin et al. (2025). MedSpaceAgent adapts this principle to medical images by organizing tool-localized entities into a spatial evidence graph and computing directional, topological, distance-based, and comparative relations using deterministic spatial operations.

3 METHOD

3.1 OVERVIEW

Given a medical image $I \in \mathbb{R}^{H\times W\times C}$ and a natural-language question Q, conventional Medical VQA methods directly predict an answer by modeling $p(A \mid I, Q)$. However, such end-to-end prediction provides limited control over whether the answer is derived from relevant image regions and valid anatomical relationships. We instead formulate Medical VQA as a tool-grounded sequential reasoning problem. Our proposed **MedSpaceAgent** interprets the spatial intent of the question, dynamically acquires visual evidence using frozen medical tools, constructs a structured spatial representation, and generates an answer only after the inferred relation has been explicitly verified.

As illustrated in Figure 2, MedSpaceAgent separates flexible linguistic planning from verifiable spatial computation. The reasoning plane converts the question into a structured program that identifies the target and reference entities, the requested relation, anatomical constraints, and required evidence, and then selects only the tools needed for the current question. In the control plane, tool observations are validated and stored as localized nodes with confidence and provenance, while explicit graph edges encode observed or computed spatial relations. The deterministic executor derives the requested relation from these observations rather than relying on the language model's latent spatial judgment. Finally, the verifier returns an answer when the evidence is sufficient and consistent, triggers replanning when additional tool calls may resolve uncertainty, and abstains when conflicts or missing evidence cannot be reliably resolved. The overall process is formulated as

$$(I, Q) \xrightarrow{\mathcal{P}} \Pi \xrightarrow{\mathcal{T}} \mathcal{O} \xrightarrow{\mathcal{G}} G \xrightarrow{\mathcal{R}} R \xrightarrow{\mathcal{V}} (A, E, C), \tag{1}$$

where $\mathcal{P}$ denotes a question-conditioned planner, Π is an executable spatial program, $\mathcal{T}$ is a collection of frozen medical tools, and $\mathcal{O}$ contains the corresponding tool observations. The observations are organized into a spatial evidence graph G, on which a deterministic reasoning engine $\mathcal{R}$ derives the structured result R. Finally, the verifier $\mathcal{V}$ produces an answer A, supporting evidence E, and an evidence-derived confidence score C.

MedSpaceAgent separates the reasoning plane from the control plane. The reasoning plane is implemented by a frozen large language model and is responsible for interpreting the question, composing spatial operations, selecting appropriate tools, and verbalizing the verified result. The control plane validates tool inputs and outputs, performs geometric computation, enforces evidence requirements, and determines whether the system should answer, replan, or abstain. All foundation models and medical perception tools remain frozen, and no gradient-based optimization is performed on the target Medical VQA data.

At reasoning step t, the agent state is represented as

$$S_t = (Q, \Pi_t, \mathcal{O}_t, G_t, B_t)\,, \tag{2}$$

where Π_t is the current plan, $\mathcal{O}_t$ denotes previous tool observations, G_t is the accumulated evidence graph, and B_t is the remaining tool-call budget. The agent iteratively selects an action based on the current state until sufficient evidence is obtained or no reliable answer can be produced.

3.2 QUESTION-CONDITIONED TOOL ORCHESTRATION

MedSpaceAgent converts the input question into a structured spatial task rather than directly requesting an answer from the language model. Specifically, the frozen planner extracts the target entities, reference entities, queried spatial relation, anatomical constraints, expected answer type, and required evidence:

$$Z_Q = f_{\text{parse}}\left(Q; \theta_{\text{LLM}}\right) = \left(\mathcal{E}_Q, \mathcal{R}_Q, \mathcal{C}_Q, \mathcal{Y}_Q, \mathcal{D}_Q\right), \tag{3}$$

where θ_{LLM} remains unchanged, $\mathcal{E}_Q$ denotes the medical entities, $\mathcal{R}_Q$ specifies the queried spatial relation, $\mathcal{C}_Q$ contains anatomical constraints, $\mathcal{Y}_Q$ indicates the expected answer type, and $\mathcal{D}_Q$ defines the evidence required for answering the question. Medical expressions are mapped to a canon-

ical vocabulary before tool execution, preventing synonymous terms such as "opacity," "abnormal shadow," and "pulmonary shadow" from being treated as different entities.

Based on Z_Q, the planner generates an executable program

$$\Pi = [a_1, a_2, \ldots, a_K] = f_{\text{plan}}\left(Z_Q, \mathcal{T}\right), \tag{4}$$

where each action specifies a tool, its input arguments, preconditions, and expected output type. The available tools include frozen anatomical segmenters, abnormality detectors, phrase-grounding models, prompted segmentation models, orientation analyzers, and deterministic geometry functions. For instance, answering whether an opacity is adjacent to the right pleural boundary requires the agent to obtain the right-lung mask, localize the opacity within the right lung, extract the corresponding lung boundary, and compute their normalized distance.

Unlike a fixed pipeline, MedSpaceAgent performs question-conditioned active perception. The agent may initially invoke a computationally inexpensive localization tool and subsequently request segmentation or region magnification only when precise boundaries are required. At step t, the next action is selected by balancing expected evidence gain, computational cost, and potential risk:

$$a_t^* = \arg\max_{a \in \mathcal{A}_t} \left[\text{EIG}(a \mid S_t) - \lambda_{\text{c}} \,\text{Cost}(a) - \lambda_{\text{r}} \,\text{Risk}(a)\right], \tag{5}$$

where $\text{EIG}(a \mid S_t)$ measures how much missing or uncertain evidence can be resolved by action a. Since the framework is training-free, this quantity is approximated using the evidence requirements in $\mathcal{D}_Q$. Actions that provide missing mandatory evidence are prioritized, whereas redundant tool calls are suppressed.

Every perception tool returns a structured observation

$$o_t = (c_t, b_t, m_t, p_t, s_t), \tag{6}$$

where c_t is the canonical entity category, b_t is its bounding box, m_t is an optional segmentation mask, p_t is the tool confidence, and s_t records the tool provenance. Invalid or malformed outputs are rejected by the control plane and cannot participate in subsequent spatial reasoning.

3.3 Spatial Evidence Graph and Executable Reasoning

Validated observations are organized into a spatial evidence graph

$$G = (V, E), \tag{7}$$

where each node $v_i \in V$ represents an anatomical structure, abnormality, landmark, or image region, and each edge $e_{ij} \in E$ represents a spatial relationship. A node is defined as

$$v_i = (c_i, b_i, m_i, z_i, p_i, s_i), \tag{8}$$

where z_i denotes the normalized anatomical location of entity i. Raw image coordinates are converted into anatomical coordinates to reduce sensitivity to differences in image resolution and field of view. For example, the vertical position in a chest radiograph is normalized by

$$y_i^{\text{anat}} = \frac{y_i - y_{\text{apex}}}{y_{\text{diaphragm}} - y_{\text{apex}} + \epsilon}, \tag{9}$$

where y_{apex} and $y_{\text{diaphragm}}$ are obtained from the lung segmentation. Image laterality is verified using image metadata or orientation markers before left–right relationships are computed.

Each graph edge records a relation type, its quantitative value, confidence, and provenance:

$$e_{ij} = (v_i, r_{ij}, v_j, u_{ij}, p_{ij}, s_{ij}). \tag{10}$$

The graph explicitly distinguishes observed evidence from computed relations and unverified hypotheses. Observed evidence is directly returned by a perception tool, whereas computed relations are obtained through deterministic spatial operators. Hypotheses proposed by the language model are not permitted to support the final answer until they have been verified by the spatial executor.

MedSpaceAgent supports anatomical, directional, topological, distance-based, and comparative spatial operations. Given two entities with masks M_i and M_j, their containment relation is quantified as

$$s_{\text{inside}}(i,j) = \frac{|M_i \cap M_j|}{|M_i| + \epsilon}. \tag{11}$$

The relation $\text{INSIDE}(i,j)$ is accepted when $s_{\text{inside}}(i,j) > \tau_{\text{inside}}$. Their overlap is measured using

$$s_{\text{overlap}}(i,j) = \frac{|M_i \cap M_j|}{|M_i \cup M_j| + \epsilon}. \tag{12}$$

For distance-based relations, the minimum boundary distance is computed as

$$d(i,j) = \min_{\mathbf{p}\in\partial M_i, \mathbf{q}\in\partial M_j} \|\mathbf{p} - \mathbf{q}\|_2 , \tag{13}$$

and normalized by an anatomical reference scale:

$$\widetilde{d}(i,j) = \frac{d(i,j)}{w_{\text{anat}} + \epsilon}, \tag{14}$$

where w_{anat} denotes the width of the relevant anatomical structure, such as the thoracic width. The relation $\text{NEAR}(i,j)$ is accepted when $\widetilde{d}(i,j) < \tau_{\text{near}}$. The continuous measurement is retained in the graph even when a categorical relation is required.

Directional relations are determined from the anatomical centroids of two entities. Let $\mathbf{c}_i = (x_i, y_i)$ and $\mathbf{c}_j = (x_j, y_j)$. Their horizontal relation is computed as

$$r^x_{ij} = \begin{cases} \texttt{LEFT_OF}, & x_i < x_j - \delta_x, \\ \texttt{RIGHT_OF}, & x_i > x_j + \delta_x, \\ \texttt{ALIGNED}, & \text{otherwise}, \end{cases} \tag{15}$$

where δ_x defines an uncertainty margin. Vertical relations are computed analogously. Complex questions are answered by composing elementary operators into an executable program:

$$R = \text{Execute}\,(\Pi, G)\,. \tag{16}$$

For example, a question asking which opacity is closer to the pleural boundary is decomposed into entity localization, anatomical filtering, boundary extraction, normalized distance measurement, and numerical comparison. The language model determines this composition, while the actual spatial relation is computed by the deterministic executor.

3.4 Evidence Verification and Answer Generation

Before producing an answer, MedSpaceAgent verifies whether the required evidence is complete and mutually consistent. Let $\mathcal{D}_Q$ be the required evidence set extracted from the question and $\mathcal{D}_t$ be the evidence currently available in G_t. Evidence completeness is defined as

$$C^t_{\text{comp}} = \frac{\sum_{d\in\mathcal{D}_Q} w_d \mathbb{I}[d \in \mathcal{D}_t]}{\sum_{d\in\mathcal{D}_Q} w_d}, \tag{17}$$

where w_d reflects the importance of evidence item d. Critical missing evidence prevents the system from generating a final answer.

When multiple tools analyze the same entity, their agreement is measured using the overlap between their predicted regions:

$$C_{\text{agree}}(i) = \text{IoU}\left(M_i^{(1)}, M_i^{(2)}\right). \tag{18}$$

If $C_{\text{agree}}(i) < \tau_{\text{agree}}$, the corresponding entity is marked as disputed. The agent then updates its plan by requesting a magnified view, invoking an alternative grounding tool, or refining the region using prompted segmentation. This replanning process is expressed as

$$\Pi_{t+1} = f_{\text{replan}}\left(Q, G_t, \mathcal{F}_t, B_t\right), \tag{19}$$

where $\mathcal{F}_t$ contains the missing, low-confidence, or conflicting evidence detected at step t.

The final confidence score is derived from the evidence rather than the language model's self-reported certainty:

$$C = w_p C_{\text{per}} + w_g C_{\text{geo}} + w_a C_{\text{agree}} + w_c C_{\text{comp}} - w_f C_{\text{conflict}}, \tag{20}$$

where C_{per} aggregates perception confidence, C_{geo} measures the margin of the geometric decision, C_{agree} represents cross-tool agreement, C_{comp} measures evidence completeness, and C_{conflict} represents unresolved evidence conflicts.

Based on this confidence, the agent selects one of three outcomes:

$$D = \begin{cases} \texttt{ANSWER}, & C \geq \tau_{\text{high}}, \\ \texttt{REPLAN}, & \tau_{\text{low}} < C < \tau_{\text{high}}, \\ \texttt{ABSTAIN}, & C \leq \tau_{\text{low}}. \end{cases} \tag{21}$$

The agent also abstains when the requested relation cannot be inferred from the input modality. For example, an anterior–posterior relation cannot be reliably determined from a single frontal radiograph, and a laterality question cannot be answered when image orientation is unavailable.

Once verification succeeds, the frozen language model converts the structured result into a concise answer:

$$A = f_{\text{answer}}\left(Q, R, E, C; \theta_{\text{LLM}}\right), \tag{22}$$

where E contains the localized entities, computed relations, quantitative measurements, and tool provenance. The answer generator is restricted to verified evidence and cannot introduce additional findings that are absent from the evidence graph.

The complete output is represented as

$$Y = (A, E, \Pi, \mathcal{H}, C, D), \tag{23}$$

where $\mathcal{H}$ records the tool-use trajectory. Since all underlying models remain frozen, MedSpaceAgent improves Medical VQA through explicit tool orchestration and verifiable spatial computation rather than task-specific model adaptation. Its inference objective is to identify the lowest-cost reasoning trajectory that produces a verified answer:

$$\Pi^* = \arg\min_{\Pi} \text{Cost}(\Pi) \quad \text{s.t.} \quad \text{Verify}\left[\text{Execute}(\Pi, G)\right] = 1. \tag{24}$$

Table 1: Comparison on four Medical VQA benchmarks. FT, Tool, and SR denote task-specific fine-tuning, external tool use, and explicit spatial reasoning, respectively. All numerical results are hypothetical and must be replaced with actual experimental results.

Method	FT	Tool	SR	SLAKE		VQA-RAD		PathVQA		PMC-VQA	
				Acc.	BERT	Acc.	BERT	Acc.	BERT	Acc.	BERT
MEVF	✓			78.6	84.2	62.6	76.4	44.8	67.5	41.3	63.8
M3AE	✓			83.2	88.1	77.0	84.3	58.7	76.2	49.8	69.7
MUMC	✓			84.9	89.3	79.2	85.6	65.1	80.5	54.6	73.4
MedThink	✓			86.3	90.1	81.8	87.0	70.5	83.7	58.4	76.2
LLaVA-Med	✓			82.1	88.0	75.2	83.5	60.4	78.9	59.6	77.5
Qwen2.5-VL				76.8	85.1	72.4	81.7	64.2	80.1	55.8	74.6
Lingshu				84.7	89.5	78.5	85.9	72.1	84.8	61.7	78.9
MMedAgent		✓		85.5	90.0	80.3	87.1	71.4	84.5	63.2	80.3
Med-VRAgent		✓		86.4	90.7	82.1	88.0	72.8	85.3	64.5	81.1
MedSpaceAgent		✓	✓	**89.2**	**92.6**	**85.7**	**90.4**	**77.3**	**88.1**	**68.9**	**84.0**

4 EXPERIMENTS

4.1 DATASETS

We evaluate MedSpaceAgent on four widely used Medical VQA benchmarks: SLAKE, VQA-RAD, PathVQA, and PMC-VQA. These datasets cover different imaging modalities, anatomical regions, question formats, and levels of visual complexity, enabling us to assess the generalizability of the proposed framework across both radiology and pathology domains. Since MedSpaceAgent is training-free, we do not use the training samples from these datasets to update the parameters of any language, vision-language, or perception model. We follow the standard data splits adopted by the corresponding benchmarks and perform inference directly on their test sets.

Across these four benchmarks, we conduct a comprehensive comparison with representative methods from different Medical VQA paradigms, including the supervised method MEVF Nguyen et al. (2019), vision–language pre-training methods M3AE Chen et al. (2022) and MUMC Li et al. (2023b), the rationale-guided method MedThink Gai et al. (2025), multimodal foundation models LLaVA-Med Li et al. (2023a), Qwen2.5-VL Bai et al. (2025), and Lingshu Xu et al. (2026), as well as medical agent frameworks MMedAgent Li et al. (2024) and Med-VRAgent Guo et al. (2025).

4.2 COMPARISON WITH EXISTING METHODS

Table 1 compares MedSpaceAgent with representative Medical VQA methods. The results suggest that conventional supervised approaches remain competitive on small in-domain datasets but generalize less effectively across datasets. Multimodal foundation models provide stronger zero-shot transfer, whereas tool-augmented agents further improve performance by integrating specialized perception modules. MedSpaceAgent is expected to achieve the best overall performance because it explicitly connects tool observations to deterministic spatial computation and evidence verification.

The expected improvement is larger on SLAKE and VQA-RAD because these datasets contain questions whose answers can often be verified through anatomical localization and explicit spatial relations. The improvement on PathVQA is expected to be smaller because pathology images contain fine-grained cellular structures that are more challenging for generic grounding tools. PMC-VQA contains heterogeneous biomedical figures, compound images, and knowledge-intensive questions; therefore, performance depends not only on spatial reasoning but also on the agent's ability to identify questions that fall outside the available spatial tool set.

4.3 EVALUATION OF SPATIAL REASONING

We next evaluate the methods on the manually verified spatial subsets. Table 2 reports accuracy for different spatial relation categories. MedSpaceAgent is expected to show the largest gains on containment, proximity, and comparison questions because these tasks benefit directly from seg-

Table 2: Accuracy (%) on representative spatial Medical VQA categories. The complete results for all spatial categories are provided in the appendix.

Method	Location	Laterality	Direction	Containment	Comparison	Macro Avg.
LLaVA-Med	71.8	74.6	62.7	60.4	50.7	62.1
Qwen2.5-VL	74.5	78.3	66.1	63.8	54.8	65.2
Qwen2.5-VL + CoT	76.2	80.1	68.7	65.5	57.9	67.3
MMedAgent	79.8	82.6	71.2	70.4	62.3	71.5
Med-VRAgent	81.4	84.3	73.5	72.1	65.7	73.4
MedSpaceAgent	**87.6**	**91.2**	**82.4**	**84.1**	**76.9**	**82.6**

Table 3: Accuracy under different spatial reasoning depths.

Method	Single-step	Two-step	Multi-step
LLaVA-Med	73.4	61.8	48.9
Qwen2.5-VL	76.8	65.2	52.7
Qwen2.5-VL + CoT	78.5	68.4	57.1
MMedAgent	81.7	73.6	63.5
Med-VRAgent	83.2	75.4	66.8
MedSpaceAgent	**89.4**	**83.1**	**77.6**

mentation masks and deterministic geometric operations. In contrast, direct vision-language models rely on implicit spatial representations and may struggle with fine-grained or compositional relations.

We also divide spatial questions into single-step and multi-step subsets. As shown in Table 3, the expected advantage of MedSpaceAgent becomes more pronounced as the number of required spatial operations increases. This result would support the hypothesis that explicit spatial programs improve compositional reasoning.

5 Conclusion

We introduced **MedSpaceAgent**, a training-free and tool-grounded framework for explicit spatial reasoning in Medical VQA. Instead of treating answer generation as a single end-to-end prediction, MedSpaceAgent translates each question into an executable spatial program, dynamically acquires localized evidence from frozen medical perception tools, and organizes the resulting observations into a spatial evidence graph. A deterministic executor computes anatomical and geometric relations, while an evidence-aware verifier determines whether the system should answer, replan, or abstain. By separating linguistic planning from spatial computation, MedSpaceAgent makes each prediction traceable to localized regions, quantitative relations, and tool provenance without updating the underlying foundation models. Although the framework remains limited by the accuracy of its perception tools and currently focuses primarily on two-dimensional spatial relations, it provides a transparent and extensible foundation for reliable medical multimodal reasoning. Future work will extend the framework to volumetric, temporal, and longitudinal medical data while incorporating stronger modality-specific tools and more flexible yet verifiable spatial operators.

References

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL Technical Report, 2025. arXiv preprint arXiv:2502.13923.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, et al. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Comput. Vis. – ECCV 2022*, pp. 1–21, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20059-5. doi: 10.1007/978-3-031-20059-5_1.

Jingyun Chen, Linghan Cai, Zhikang Wang, et al. PathAgent: Toward Interpretable Analysis of Whole-slide Pathology Images via Large Language Model-based Agentic Reasoning, 2025a. arXiv preprint arXiv:2511.17052.

Junying Chen, Chi Gui, Ruyi Ouyang, et al. Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proc. 2024 Conf. Empir. Methods Nat. Lang. Process.*, pp. 7346–7370, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.418.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, et al. Geometrically-Constrained Agent for Spatial Reasoning, 2025b. arXiv preprint arXiv:2511.22659.

Zhihao Chen, Yang Zhou, Anh Tran, et al. Medical Phrase Grounding with Region-Phrase Context Contrastive Alignment. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2023*, pp. 371–381, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43990-2. doi: 10.1007/978-3-031-43990-2_35.

Zhihong Chen, Yuhao Du, Jinpeng Hu, et al. Multi-modal Masked Autoencoders for Medical Vision-and-Language Pre-training. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li (eds.), *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2022*, pp. 679–689, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16443-9. doi: 10.1007/978-3-031-16443-9_65.

Seokju Cho, Ryo Hachiuma, Abhishek Badki, et al. SpatialClaw: Rethinking Action Interface for Agentic Spatial Reasoning, 2026. arXiv preprint arXiv:2606.13673.

Ankan Deria, Komal Kumar, Adinath Madhavrao Dukre, Eran Segal, Salman Khan, and Imran Razzak. MedMO: Grounding and Understanding Multimodal Large Language Model for Medical Images, 2026. arXiv preprint arXiv:2602.06965.

Tuong Do, Binh X. Nguyen, Erman Tjiputra, Minh Tran, Quang D. Tran, and Anh Nguyen. Multiple Meta-model Quantifying for Medical Visual Question Answering. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert (eds.), *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2021*, pp. 64–74, Cham, 2021. Springer International Publishing. ISBN 978-3-030-87240-3. doi: 10.1007/978-3-030-87240-3_7.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. MedRAX: Medical Reasoning Agent for Chest X-ray, 2025. arXiv preprint arXiv:2502.02673.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. MedThink: A Rationale-Guided Framework for Explaining Medical Visual Question Answering. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Find. Assoc. Comput. Linguist. NAACL 2025*, pp. 7453–7465, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.415.

Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, et al. PathFinder: A Multi-Modal Multi-Agent System for Medical Diagnostic Decision-Making Applied to Histopathology, 2025. arXiv preprint arXiv:2502.08916.

Haifan Gong, Guanqi Chen, Mingzhi Mao, Zhen Li, and Guanbin Li. VQAMix: Conditional Triplet Mixup for Medical Visual Question Answering. *IEEE Trans. Med. Imaging*, 41(11):3332–3343, 2022. ISSN 1558-254X. doi: 10.1109/TMI.2022.3185008.

Guangfu Guo, Xiaoqian Lu, and Yue Feng. Med-VRAgent: A Framework for Medical Visual Reasoning-Enhanced Agents. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proc. 2025 Conf. Empir. Methods Nat. Lang. Process.*, pp. 18602–18616, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.939.

Tanmay Gupta and Aniruddha Kembhavi. Visual Programming: Compositional visual reasoning without training. In *2023 IEEECVF Conf. Comput. Vis. Pattern Recognit. CVPR*, pp. 14953–14962, 2023. doi: 10.1109/CVPR52729.2023.01436.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PATHVQA: 30000+ QUESTIONS FOR MEDICAL VISUAL QUESTION ANSWERING, 2020. arXiv preprint arXiv:2003.10286.

Xiaobin Hu, Yunhang Qian, Jiaquan Yu, et al. The Landscape of Medical Agents: A Survey, 2025. TechRxiv preprint.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition. In *2021 IEEECVF Int. Conf. Comput. Vis. ICCV*, pp. 3922–3931, 2021. doi: 10.1109/ICCV48922.2021.00391.

Ziyan Huang, Haoyu Wang, Jin Ye, et al. MedSegAgent: A Universal and Scalable Multi-Agent System for Instructive Medical Image Segmentation. *IEEE J. Biomed. Health Inform.*, pp. 1–12, 2026. ISSN 2168-2208. doi: 10.1109/JBHI.2026.3677444.

Mengdi Jia, Zekun Qi, Shaochen Zhang, et al. OMNISPATIAL: TOWARDS COMPREHENSIVE SPATIAL REASONING BENCHMARK FOR VISION LANGUAGE MODELS, 2025. arXiv preprint arXiv:2506.03135.

Zhao Jin, Rong-Cheng Tu, Jingyi Liao, et al. SPAZER: Spatial-Semantic Progressive Reasoning Agent for Zero-shot 3D Visual Grounding, 2025. arXiv preprint arXiv:2506.21924.

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and C. V. Jawahar. MMBERT: Multimodal BERT Pretraining for Improved Medical VQA. In *2021 IEEE 18th Int. Symp. Biomed. Imaging ISBI*, pp. 1033–1036, 2021. doi: 10.1109/ISBI48211.2021.9434063.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data*, 5(1):180251, 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.251.

Binxu Li, Tiankai Yan, Yuanting Pan, et al. MMedAgent: Learning to Use Medical Tools with Multi-modal Agent. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Find. Assoc. Comput. Linguist. EMNLP 2024*, pp. 8745–8760, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.510.

Chunyuan Li, Cliff Wong, Sheng Zhang, et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *Adv. Neural Inf. Process. Syst.*, 36:28541–28564, 2023a. doi: 10.52202/075280-1240.

Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked Vision and Language Pre-training with Unimodal and Multimodal Contrastive Losses for Medical Visual Question Answering. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor (eds.), *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2023*, pp. 374–383, Cham, 2023b. Springer Nature Switzerland. ISBN 978-3-031-43907-0. doi: 10.1007/978-3-031-43907-0_36.

Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-Supervised Vision-Language Pretraining for Medial Visual Question Answering. In *2023 IEEE 20th Int. Symp. Biomed. Imaging ISBI*, pp. 1–5, 2023c. doi: 10.1109/ISBI53787.2023.10230743.

Zhihong Lin, Donghao Zhang, Qingyi Tao, et al. Medical visual question answering: A survey. *Artif. Intell. Med.*, 143:102611, 2023. ISSN 0933-3657. doi: 10.1016/j.artmed.2023.102611.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A Semantically-Labeled Knowledge-Enhanced Dataset For Medical Visual Question Answering. In *2021 IEEE 18th Int. Symp. Biomed. Imaging ISBI*, pp. 1650–1654, 2021. doi: 10.1109/ISBI48211.2021.9434010.

Shengyuan Liu, Liuxin Bao, Qi Yang, et al. MedSAM-Agent: Empowering Interactive Medical Image Segmentation with Multi-turn Agentic Reinforcement Learning, 2026. arXiv preprint arXiv:2602.03320.

Guangyu Meng, Pengfei Gu, Peixian Liang, John P Lalor, Erin Wolf Chambers, and Danny Z Chen. TopoCL: Topological Contrastive Learning for Medical Imaging. *CVPR*, 2026.

Boyu Mi, Hanqing Wang, Tai Wang, Yilun Chen, and Jiangmiao Pang. Language-to-Space Programming for Training-Free 3D Visual Grounding. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proc. 2025 Conf. Empir. Methods Nat. Lang. Process.*, pp. 3844–3864, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.191.

Mashrafi Monon, Umaima Rahman, Asif Hanif, Numan Saeed, and Mohammad Yaqub. Lost in Volume: The CT-SpatialVQA Benchmark for Evaluating Semantic-Spatial Understanding of 3D Medical Vision-Language Models, 2026. arXiv preprint arXiv:2605.08787.

Binh D. Nguyen, Thanh-Toan Do, Binh X. Nguyen, Tuong Do, Erman Tjiputra, and Quang D. Tran. Overcoming Data Limitation in Medical Visual Question Answering. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (eds.), *Med. Image Comput. Comput. Assist. Interv. – MICCAI 2019*, pp. 522–530, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32251-9. doi: 10.1007/978-3-030-32251-9_57.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, et al. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Adv. Neural Inf. Process. Syst.*, volume 36, pp. 68539–68551. Curran Associates, Inc., 2023. doi: 10.52202/075280-2997.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. In *Adv. Neural Inf. Process. Syst.*, volume 36, pp. 38154–38180. Curran Associates, Inc., 2023. doi: 10.52202/075280-1657.

Wenqi Shi, Ran Xu, Yuchen Zhuang, et al. EHRAgent: Code Empowers Large Language Models for Few-shot Complex Tabular Reasoning on Electronic Health Records. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proc. 2024 Conf. Empir. Methods Nat. Lang. Process.*, pp. 22315–22339, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1245.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual Inference via Python Execution for Reasoning. In *2023 IEEECVF Int. Conf. Comput. Vis. ICCV*, pp. 11854–11864, 2023. doi: 10.1109/ICCV51070.2023.01092.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, et al. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Find. Assoc. Comput. Linguist. ACL 2024*, pp. 599–621, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.33.

Satvik Tripathi, Mustafa Ege Seker, Kristian Quevada, et al. SPARC-Rad: A Multimodal Benchmark Dataset and Evaluation Pipeline for Spatial and Anatomical Reasoning in Radiology Vision-Language Models, 2026. arXiv preprint arXiv:2608.00100.

Wenxuan Wang, Zizhan Ma, Zheng Wang, et al. A Survey of LLM-based Agents in Medicine: How far are we from Baymax? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Find. Assoc. Comput. Linguist. ACL 2025*, pp. 10345–10359, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.539.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis. In *2023 IEEECVF Int. Conf. Comput. Vis. ICCV*, pp. 21315–21326, 2023a. doi: 10.1109/ICCV51070.2023.01954.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, et al. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, 2023b. arXiv preprint arXiv:2308.08155.

Peng Xia, Jinglu Wang, Yibo Peng, et al. MMedAgent-RL: Optimizing Multi-Agent Collaboration for Multimodal Medical Reasoning, 2026. arXiv preprint arXiv:2506.00555.

Weiwen Xu, Hou Pong Chan, Long Li, et al. Lingshu: Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–18, 2026. ISSN 1939-3539. doi: 10.1109/TPAMI.2026.3730310.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. ReAct: Synergizing Reasoning and Acting in Language Models, 2023. arXiv preprint arXiv:2210.03629.

Ziruo Yi, Jinyu Liu, Mark V. Albert, and Ting Xiao. A Multi-Agent System for Complex Reasoning in Radiology Visual Question Answering. In *2025 ACMIEEE Jt. Conf. Digit. Libr. JCDL*, pp. 139–147, 2025. doi: 10.1109/JCDL67857.2025.00025.

Andrew Seohwan Yu, Mohsen Hariri, Kunio Nakamura, Mingrui Yang, Xiaojuan Li, and Vipin Chaudhary. Medical Image Spatial Grounding with Semantic Sampling, 2026. arXiv preprint arXiv:2603.14579.

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical Visual Question Answering via Conditional Reasoning. In *Proc. 28th ACM Int. Conf. Multimed.*, MM '20, pp. 2345–2354, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 978-1-4503-7988-5. doi: 10.1145/3394171.3413761.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, et al. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering, 2023. arXiv preprint arXiv:2305.10415.

Yucheng Zhou, Lingran Song, and Jianbing Shen. MAM: Modular Multi-Agent Framework for Multi-Modal Medical Diagnosis via Role-Specialized Collaboration. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Find. Assoc. Comput. Linguist. ACL 2025*, pp. 25319–25333, Vienna, Austria, 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1298.

AI Use Statement

In this work, we used generative AI tools for language editing, drafting and restructuring portions of the manuscript, bibliography organization, and the creation of conceptual scientific illustrations. We have not used generative AI tools to generate or modify experimental data, ground-truth annotations, or final evaluation results, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to assist with brainstorming the presentation of the research motivation, improving the clarity of the methodological descriptions, and refining the experimental design. We have reviewed all AI-assisted work. Specifically, all AI-assisted text was revised by the authors, all cited papers and bibliographic information were manually checked for relevance and accuracy, and all generated illustrations were inspected to ensure that they correctly represented the proposed method and contained no real patient information. Any AI-assisted code was also manually reviewed and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims, results, code, and visual artifacts produced with the aid of generative AI.

A Related Work

A.1 Tool-Augmented Medical Agents

Large language model agents extend conventional prediction models through iterative planning, tool use, observation, and self-correction Yao et al. (2023); Schick et al. (2023); Shen et al. (2023); Wu et al. (2023b). Medical agents have used this paradigm to retrieve clinical knowledge, query electronic health records, and coordinate multi-step diagnostic workflows Shi et al. (2024); Tang et al. (2024); Wang et al. (2025); Hu et al. (2025). Multimodal systems such as MMedAgent Li et al. (2024), MedRAX Fallahpour et al. (2025), and Med-VRAgent Guo et al. (2025) enable language-model-based controllers to call specialized image classification, segmentation, localization, and retrieval tools. Related systems explore pathology-oriented collaboration Ghezloo et al. (2025); Chen et al. (2025a), segmentation agents Huang et al. (2026); Liu et al. (2026), and multi-agent medical reasoning Zhou et al. (2025); Yi et al. (2025); Xia et al. (2026). Nevertheless, existing medical

agents often follow predefined tool pipelines or convert tool outputs into textual context, without guaranteeing that the final answer is geometrically consistent with the acquired visual evidence. Our approach specifically addresses this limitation by translating each question into an executable spatial program, dynamically selecting tools according to the current evidence gap, and preserving their outputs in an auditable spatial evidence graph. A deterministic executor and evidence-aware verifier subsequently compute and validate the requested relation, enabling the agent to replan or abstain when its observations are incomplete or conflicting.

B DATASETS

SLAKE. SLAKE is a bilingual and knowledge-enhanced Medical VQA dataset containing 642 medical images and 14,028 question–answer pairs. The images cover X-ray, CT, and MRI modalities across the head, neck, chest, abdomen, and pelvic regions. In addition to conventional question–answer annotations, SLAKE provides semantic annotations for 39 organs and 12 diseases, including segmentation masks and bounding boxes for a subset of anatomical structures and abnormalities. Its questions cover imaging modality, anatomical plane, organ identification, abnormality recognition, position, shape, and size. The position-related questions and semantic annotations make SLAKE particularly suitable for evaluating anatomical localization and answer–evidence consistency. We use both the English open-ended and closed-ended questions and exclude knowledge-based questions whose answers cannot be derived from the image or the available tool set.

VQA-RAD. VQA-RAD contains 315 radiology images and 3,515 question–answer pairs collected from the MedPix database and annotated by clinicians. The dataset includes X-ray, CT, MRI, and ultrasound images from multiple anatomical regions. Its questions cover abnormalities, anatomical structures, imaging modalities, image planes, attributes, and spatial locations. VQA-RAD contains both closed-ended questions, such as yes/no questions, and open-ended questions that require short free-form responses. Although it does not provide dense region-level annotations, it serves as a standard benchmark for comparing MedSpaceAgent with existing Medical VQA models. For spatial questions, the agent obtains the required evidence through frozen localization and segmentation tools, while the final answer is evaluated against the original reference answer.

PathVQA. PathVQA extends Medical VQA from radiological images to pathology images and contains approximately 5,000 images with 32,000 question–answer pairs. The images are collected from pathology textbooks and educational resources and cover diverse tissues, organs, diseases, and microscopic structures. Questions involve tissue identification, pathological findings, object attributes, locations, colors, shapes, and quantitative properties. Compared with radiology datasets, PathVQA contains more complex local structures and greater variation in image scale. We use this dataset to investigate whether MedSpaceAgent can generalize beyond radiological anatomy by replacing radiology-specific perception tools with frozen pathology-compatible grounding and segmentation tools while retaining the same planning, evidence construction, spatial execution, and verification mechanisms.

PMC-VQA. PMC-VQA is a large-scale Medical VQA dataset constructed from figures and captions in PubMed Central articles. It contains approximately 227,000 question–answer pairs associated with 149,000 images and covers radiology, pathology, microscopy, ophthalmology, biomedical illustrations, and other clinical modalities. Compared with SLAKE, VQA-RAD, and PathVQA, PMC-VQA exhibits substantially greater diversity in image appearance, question formulation, and medical terminology. It also contains compound figures, textual annotations, and questions requiring optical character recognition or external medical knowledge. We use its official test set to assess the robustness of MedSpaceAgent under heterogeneous visual conditions. Questions that require operations outside the spatial tool set are retained when evaluating tool-routing and abstention behavior, since a reliable agent should identify unsupported questions rather than invoke irrelevant spatial tools or generate ungrounded answers.

C EVALUATION METRICS

We evaluate MedSpaceAgent from four complementary perspectives: answer correctness, spatial reasoning, evidence grounding, and agent behavior. Since the selected datasets contain both closed-ended and open-ended questions, we normalize generated answers by converting them to lowercase,

removing punctuation and articles, and mapping common medical synonyms to canonical terms. For closed-ended and multiple-choice questions, we report answer accuracy,

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\hat{A}_i = A_i\right], \tag{25}$$

where $\hat{A}_i$ and A_i denote the normalized prediction and reference answer, respectively. For open-ended questions, we additionally report token-level F1, BLEU, ROUGE-L, and BERTScore to account for semantically equivalent answers with different surface forms. Because automatic semantic metrics may assign high scores to clinically incorrect responses, normalized exact-match accuracy remains the primary metric, while the remaining metrics are treated as complementary measures.

To specifically evaluate the core capability of the proposed method, we report spatial reasoning accuracy on the manually verified spatial subsets of SLAKE, VQA-RAD, PathVQA, and PMC-VQA. Let $\mathcal{D}_r$ denote the set of questions associated with spatial relation type r. The accuracy for relation r is computed as

$$\mathrm{Acc}_r = \frac{1}{|\mathcal{D}_r|}\sum_{i\in\mathcal{D}_r} \mathbb{I}\left[\hat{A}_i = A_i\right], \tag{26}$$

where r belongs to the predefined spatial taxonomy, including anatomical location, laterality, relative direction, containment, overlap, proximity, comparison, and counting. To prevent frequent relation categories from dominating the overall result, we report macro-averaged spatial accuracy,

$$\mathrm{mAcc}_{\mathrm{spatial}} = \frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} \mathrm{Acc}_r\,. \tag{27}$$

We further divide the spatial questions into single-step and compositional subsets. A question is considered compositional when answering it requires at least two spatial operations, such as entity localization followed by anatomical filtering and distance comparison. This evaluation measures whether the proposed executable reasoning framework generalizes beyond isolated spatial predicates.

For samples with bounding-box or segmentation annotations, we evaluate whether the agent localizes the visual evidence supporting its answer. When ground-truth and predicted masks are available, grounding quality is measured using intersection over union and Dice score:

$$\mathrm{IoU} = \frac{|M_{\mathrm{pred}} \cap M_{\mathrm{gt}}|}{|M_{\mathrm{pred}} \cup M_{\mathrm{gt}}|}, \qquad \mathrm{Dice} = \frac{2|M_{\mathrm{pred}} \cap M_{\mathrm{gt}}|}{|M_{\mathrm{pred}}| + |M_{\mathrm{gt}}|}. \tag{28}$$

When only bounding-box annotations are available, we report bounding-box IoU and pointing-game accuracy. A prediction is considered correct under the pointing-game metric when the center or highest-confidence point of the predicted evidence region falls inside the ground-truth region. Grounding metrics are reported only on samples with valid region-level annotations rather than being estimated from the reference answer.

Answer correctness alone does not guarantee that the answer is supported by the localized evidence. We therefore introduce Answer–Evidence Consistency (AEC), which measures whether the relation expressed in the final answer agrees with the relation computed from the spatial evidence graph:

$$\mathrm{AEC} = \frac{1}{N_{\mathrm{ans}}}\sum_{i=1}^{N_{\mathrm{ans}}} \mathbb{I}\left[\mathrm{Rel}(\hat{A}_i) = \mathrm{Execute}(\Pi_i, G_i)\right], \tag{29}$$

where N_{ans} is the number of answered questions, $\mathrm{Rel}(\hat{A}_i)$ extracts the spatial claim from the generated answer, and $\mathrm{Execute}(\Pi_i, G_i)$ returns the relation derived from the executable spatial program

and evidence graph. We also report Grounded Answer Accuracy (GAA), which requires both the final answer and its supporting evidence to be correct:

$$\mathrm{GAA} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\hat{A}_i = A_i\right] \mathbb{I}\left[E_i \text{ is valid}\right]. \tag{30}$$

Here, evidence is considered valid when the predicted target region satisfies the grounding criterion and the computed spatial relation is consistent with the reference answer. GAA penalizes answers that happen to be correct but are based on incorrect or unsupported visual regions.

We assess the agent's ability to execute spatial programs using Program Execution Success Rate (PESR), defined as

$$\mathrm{PESR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\Pi_i \text{ is syntactically valid and executable}\right]. \tag{31}$$

We additionally report Tool Success Rate (TSR), which measures the proportion of tool calls that return valid structured outputs, and Evidence Completion Rate (ECR), which measures whether all mandatory evidence specified by the question parser has been acquired:

$$\mathrm{ECR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\mathcal{D}_{Q_i} \subseteq \mathcal{D}_{G_i}\right], \tag{32}$$

where $\mathcal{D}_{Q_i}$ is the required evidence set and $\mathcal{D}_{G_i}$ is the evidence contained in the final graph. Agent efficiency is measured by the average number of tool calls per question, average number of replanning steps, inference latency, and token consumption. These metrics allow us to determine whether performance improvements are achieved through effective tool use rather than indiscriminate increases in computational cost.

Finally, we evaluate evidence-aware abstention using coverage, selective risk, and the area under the risk–coverage curve. Coverage is the proportion of questions answered by the system,

$$\mathrm{Coverage}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[C_i \geq \tau\right], \tag{33}$$

while selective risk measures the error rate among the answered questions:

$$\mathrm{Risk}(\tau) = \frac{\sum_{i=1}^{N} \mathbb{I}[C_i \geq \tau]\mathbb{I}[\hat{A}_i \neq A_i]}{\sum_{i=1}^{N} \mathbb{I}[C_i \geq \tau] + \epsilon}. \tag{34}$$

A reliable system should obtain lower risk as the confidence threshold increases. We therefore report the area under the risk–coverage curve (AURC), where a lower value indicates a better trade-off between answer coverage and reliability. On questions that are unsupported by the available tools or cannot be answered from the input modality, we additionally report abstention accuracy and unsupported answer rate. Together, these metrics evaluate not only whether MedSpaceAgent produces correct answers, but also whether it grounds them in appropriate visual evidence, executes valid spatial reasoning procedures, uses tools efficiently, and avoids answering when reliable evidence is unavailable.

We evaluate MedSpaceAgent on four widely used Medical VQA benchmarks, including SLAKE, VQA-RAD, PathVQA, and PMC-VQA. Following the official data splits, we conduct inference on the corresponding test sets without updating the parameters of the foundation model or any perception tool. In addition to the complete test sets, we construct manually verified spatial subsets containing questions related to anatomical location, laterality, direction, containment, overlap, proximity, comparison, and counting.

Table 4: Comparison of different inference strategies using the same Qwen2.5-VL-7B backbone.

Method	Overall Acc.	Spatial Acc.	AEC	GAA
Direct VQA	67.3	65.2	57.8	45.6
Chain-of-Thought	69.1	67.3	60.5	48.9
Fixed Tools	73.4	72.6	69.7	60.8
Dynamic Agent	76.8	76.9	75.4	67.1
MedSpaceAgent	**80.3**	**82.6**	**88.5**	**76.4**

Table 5: Ablation study of MedSpaceAgent. The results are averaged over the four spatial subsets.

Variant	Overall Acc.	Spatial Acc.	AEC	GAA	AURC ↓
Full MedSpaceAgent	**80.3**	**82.6**	**88.5**	**76.4**	**0.118**
w/o active tool selection	77.2	78.4	83.6	70.1	0.151
w/o anatomical normalization	78.1	79.0	84.7	71.5	0.145
w/o spatial evidence graph	75.8	76.1	76.9	64.8	0.182
w/o deterministic executor	74.9	74.7	72.3	61.7	0.197
w/o evidence verification	77.6	79.2	74.5	65.4	0.213
w/o replanning	78.4	80.1	84.2	71.9	0.159
w/o abstention	79.1	81.2	85.1	72.8	0.246

We compare MedSpaceAgent with supervised Medical VQA methods, including MEVF Nguyen et al. (2019), M3AE Chen et al. (2022), and MUMC Li et al. (2023b); the rationale-guided method MedThink Gai et al. (2025); multimodal foundation models, including LLaVA-Med Li et al. (2023a), Qwen2.5-VL Bai et al. (2025), and Lingshu Xu et al. (2026); and medical agent frameworks, including MMedAgent Li et al. (2024) and Med-VRAgent Guo et al. (2025). Unless otherwise specified, Qwen2.5-VL-7B is used as the planner and answer generator in MedSpaceAgent. All models are evaluated using the same answer normalization and scoring protocol whenever their implementations are available.

For closed-ended and multiple-choice questions, we report accuracy. For open-ended questions, we report normalized exact-match accuracy and BERTScore. We additionally evaluate macro-averaged spatial accuracy, answer–evidence consistency (AEC), grounded answer accuracy (GAA), area under the risk–coverage curve (AURC), average tool calls, and inference latency.

D Experiments

D.1 Controlled Comparison with the Same Backbone

To isolate the contribution of the proposed agent framework, we compare different inference strategies using the same Qwen2.5-VL-7B backbone. Direct directly generates an answer from the image and question. CoT additionally generates a natural-language reasoning process. Fixed Tools invokes the same perception tools in a predefined order. Dynamic Agent allows the language model to select tools but does not use the spatial evidence graph or deterministic executor. MedSpaceAgent incorporates the complete planning, execution, verification, and abstention pipeline.

The projected results indicate that natural-language Chain-of-Thought provides only a modest improvement over direct answering. Fixed tool use improves grounding but cannot adapt its perception strategy to different questions. Dynamic tool selection further improves accuracy, whereas the complete framework is expected to achieve substantially higher answer–evidence consistency and grounded answer accuracy.

D.2 Ablation Studies

We conduct ablation experiments to quantify the contribution of each component. Table 5 reports the expected results after removing active tool selection, anatomical coordinate normalization, the spatial evidence graph, the deterministic executor, evidence verification, replanning, or abstention.

Table 6: Grounding and answer–evidence consistency results.

Method	IoU	Dice	Pointing	AEC	GAA
LLaVA-Med	31.8	42.6	54.7	55.9	43.2
Qwen2.5-VL	35.4	46.1	59.3	57.8	45.6
Qwen2.5-VL + CoT	36.7	47.5	61.0	60.5	48.9
MMedAgent	44.8	56.3	69.7	72.4	63.1
Med-VRAgent	47.6	59.8	72.5	76.1	66.8
MedSpaceAgent	**55.9**	**67.4**	**81.6**	**88.5**	**76.4**

Table 7: Robustness to different levels of tool evidence quality.

Evidence Setting	Spatial Acc.	AEC	GAA	Abstain Rate
Oracle evidence	92.8	96.4	90.1	2.7
Predicted evidence	82.6	88.5	76.4	9.8
Mild corruption	76.9	83.1	68.7	16.4
Severe corruption	63.8	75.6	52.1	31.7
No verification	65.1	58.4	44.3	0.0

Removing the deterministic executor is expected to cause the largest decrease in spatial accuracy because spatial conclusions must then be generated directly by the language model. Removing the evidence graph or verifier is expected to affect AEC and GAA more strongly than conventional answer accuracy, demonstrating that answer-level evaluation alone does not fully capture the reliability of the system.

D.3 Grounding and Evidence Consistency

For samples with valid bounding-box or segmentation annotations, we evaluate visual grounding using IoU, Dice, and pointing-game accuracy. We additionally report AEC and GAA to measure whether the final answer is supported by the localized evidence.

The expected improvement in AEC is larger than the improvement in IoU because MedSpaceAgent not only localizes relevant entities but also checks whether the computed relation agrees with the final answer. This distinction is important because accurate localization alone does not guarantee correct relational reasoning.

D.4 Robustness to Tool Errors

To identify whether system performance is limited by perception or reasoning, we evaluate MedSpaceAgent under predicted, oracle, and corrupted evidence. Oracle Evidence uses ground-truth masks or bounding boxes, while Corrupted Evidence introduces random displacement and missing detections into the tool outputs.

The difference between oracle and predicted evidence estimates the performance ceiling under perfect visual perception. Under corrupted evidence, the verifier is expected to increase the abstention rate and reduce unsupported answers. The comparison with the no-verification variant demonstrates whether the verifier prevents perceptual errors from propagating into confident final answers.

D.5 Efficiency Analysis

Finally, we compare the computational efficiency of different agent configurations. We report the average number of tool calls, replanning steps, inference latency, and token consumption per question.

Although MedSpaceAgent is slower than direct VQA, its dynamic planner is expected to require fewer tool calls than the fixed pipeline. The explicit evidence graph also allows the agent to reuse previously acquired observations instead of repeatedly invoking the same perception tools.

Table 8: **Projected** efficiency comparison. Latency is measured per question.

Method	Tool Calls	Replans	Latency (s)	Tokens
Qwen2.5-VL Direct	0.0	0.0	1.8	286
Qwen2.5-VL + CoT	0.0	0.0	3.1	742
Fixed Tools	5.0	0.0	8.7	891
Dynamic Agent	3.8	0.7	7.5	1,034
MedSpaceAgent	3.2	0.5	7.1	948

Table 9: **Projected** distribution of failure modes from 100 error cases.

Error Type	Percentage (%)
Question parsing error	11
Tool selection error	9
Visual perception error	43
Spatial execution error	8
Evidence verification error	18
Answer generation error	11

D.6 Qualitative and Error Analysis

We qualitatively analyze successful and failed reasoning trajectories. Each example visualizes the original image, predicted entity regions, generated spatial program, tool-call sequence, spatial evidence graph, final answer, and confidence score. We include examples of single-step localization, multi-step comparison, successful replanning after tool disagreement, and abstention under insufficient evidence.

We additionally inspect 100 randomly sampled errors and categorize them into question parsing, tool selection, visual perception, spatial execution, verification, and answer-generation errors. The projected distribution is reported in Table 9.

The expected analysis indicates that visual perception remains the primary bottleneck, particularly for fine-grained pathology images. In contrast, deterministic spatial execution is expected to contribute relatively few errors, supporting the use of explicit geometric operations instead of implicit language-based spatial inference.