Interpretable Medical Vision-Language Inference with Spatial-Temporal Anatomical Graphs

Anonymous ACL submission

Abstract

Interpretable Medical cross-modal reasoning remains challenging and leads to opaque, uncontrollable, and often hallucinated outputs. To address these issues, we propose a framework, which introduces explicit and interpretable reasoning into medical vision-language tasks. First, we construct a **Spatial-Temporal Anatomical Graph (STAG)** by combining detection models with a visual-prompt-guided VLM, where nodes represent anatomical structures and edges capture *hierarchical*, *spatial*, and *temporal* relations. Then, we design the **Thinking-with-Graph (TwG)** paradigm to reason over the STAG through bottom-up traversal, integrating local evidence into global conclusions while considering alternative diagnostic paths. This design produces a transparent and controllable *thinking process*, enabling clinicians to inspect intermediate steps and better understand the model's decisions. Experiments on large-scale chest X-ray datasets show that our method achieves performance comparable to state-of-the-art baselines while providing significantly improved interpretability.

1 Introduction

In clinical practice, radiologists and physicians typically interpret medical images based on the findings for each anatomical structure. They integrate findings across different structure with patients' longitudinal data to construct diagnostic understanding and decision—a process that remains difficult to replicate with automated methods. However, given the scarcity of medical resources, the repetitive nature of report writing consumes a substantial portion of clinicians' time. To alleviate this burden, recent research has explored leveraging medical foundation vision–language models (VLMs) to assist in report generation and other cross-modal medical tasks. Existing works, represented by foundational medical-domain VLMs (Ferber et al., 2024; Zhu et al., 2025; Moor et al., 2023; Huang et al., 2024) such as LLaVA-Med (Li et al., 2024), primarily focus on aligning medical images with textual representations, aiming to improve language models' understanding of medical images and the quality of downstream generation tasks. Current works typically take a single image as input and produce a whole-image diagnosis or report through an opaque, black-box process. Such end-to-end prediction makes the decision-making pathway unclear: intermediate reasoning steps remain hidden, and local pieces of evidence cannot be explicitly enumerated. As a result, clinicians often find that the system's outputs are "correct without clarity" or "wrong without explanation," limiting both trust and adoption. In conclusion, current VLM driven multi-modal understanding frameworks still face critical limitations: they often lack interpretability and fail to emulate the realistic workflows resonating with medical practitioners' thinking process.

Building on the foundational VLM models, some follow-up studies (Kumar et al., 2025; Ali, 2025) attempt to enhance interpretability by either employing a post-hoc module or developing the intrinsic explainability. Some works rely on post-hoc tool such as Grad-CAM (Selvaraju et al., 2019) to visualize the attention heatmaps of the VLMs. However, such visualizations still provide no explicit evidence explaining why the model generates a particular diagnostic statement or report. Another line of research, exemplified by RGRG (Tanida et al., 2023) and AP-ISG (Zhang et al., 2024a), aims to enhance intrinsic interpretability by decomposing images into anatomical structures and generating fine-grained descriptions for each region. They fail to capture the logical sequence of evidence integration, resulting in explanations that are coarse and insufficient for medical decision-making. Therefore, to accelerate the development of medical decision-support models, it is imperative to devise an approach that enhances

the interpretability of large medical models' reasoning while emulating real physicians' decision-making processes and effectively leveraging evidence across different anatomical structures. Physicians' diagnostic reasoning is inherently non-linear, often resembling graph-based processes. They integrate spatial correlations across anatomical structures, incorporate longitudinal information from prior reports, and weigh evidence at multiple hierarchical levels, from global findings down to sub-regional details. Although these studies enhance interpretability by first generating structure-level findings they do not address how to organize these local discoveries into a coherent reasoning process—analogous to a chain-of-thought—that enables the model to construct holistic and globally consistent observations. Therefore, we want to use a framework to reflect how physicians organize and update knowledge, enabling models to emulate non-linear diagnostic thinking and generate clinically trustworthy reports.

To be specific, we introduce a two-stage framework designed to incorporate fine-grained structure-level reasoning into medical image understanding. In the first stage, we construct a **Spatial-Temporal Anatomical Graph (STAG)** by leveraging a visual-prompt-guided VLM (Zhu et al., 2025; Tanida et al., 2023) together with a detection model. Each vertex of the STAG corresponds to an anatomical structure within the chest X-ray—or its retrieved historical version from prior examinations—while the edges encode three types of relations: *temporal*, *spatial*, and *hierarchical*. This representation grounds the analysis in clinically meaningful anatomy and enables modeling of both compositional and longitudinal dependencies. To ensure that the reasoning process not only aggregates fine-grained details but also produces coherent conclusions at the global level, we adopt a **bottom-up traversal** strategy. This design allows information from local anatomical regions to be progressively integrated into higher-level structures, mirroring the diagnostic workflow of physicians. Moreover, to enrich the reasoning process, we aim to emulate how clinicians consider alternative diagnostic paths by selectively visiting different neighboring nodes and applying operators that filter, rank, and aggregate evidence. Building on this motivation, in the second stage we propose the **Thinking-with-Graph (TwG)** paradigm to perform structured reasoning over the constructed STAG. The TwG algorithm defines a bottom-up execution order via a controller, determines which nodes to expand through a planner, integrates textual evidence sequentially into a dynamic memory bank, and refines findings across neighboring nodes using ranking and aggregation operators. Our framework is compatible with various downstream tasks, including the most common medical VLM tasks such as VQA or Medical Report Generation(MRG). Therefore, depending on the downstream task, TwG can either traverse the entire graph to generate comprehensive medical reports or selectively expand subgraphs for focused question answering. Crucially, all intermediate reasoning steps are transparent and editable, allowing clinicians to examine and adjust the process before producing the final output.

2 Related Work

2.1 Medical Visual Grounding Models

Object grounding and grounding is a fine-grained recognizing visual task, which is challenging, particularly in the medical domain. Recently, some pioneering works (Qin et al., 2023; Zhu et al., 2025; Yang et al., 2025) aim for utilizing VLMs to guide the visual grounding task in medical domain. In Grounding-DINO (Liu et al., 2023), the model can take language prompt to locate corresponding objects. In MVLP (Qin et al., 2023) and Struct-GLIP (Zhang et al., 2024b), researchers leverage the domian knowledge in language to enhance the generalization capability of pre-trianed large Visual Grounding models.

2.2 Vision Language Models and MRG

Recently, Vision Language Models (Li et al., 2023; Liu et al., 2024) have been developed rapidly. Following this line of work, some researchers developed Medical-Domain specialized VLMs (Moor et al., 2023; Lin et al., 2023; Wu et al., 2023), which are pre-trained on large medical image and text report pairwise datasets (Johnson et al., 2019; Zhang et al., 2023; Chen et al., 2024a). Except for these large-scale pre-trained models in medical domain, many recent works have been focused on downstream tasks such as Medical VQA (Zhang et al., 2023; Zhu et al., 2025) and medical report generation (Wu et al., 2023; Huang et al., 2023, 2024; Gu et al., 2023; Chen et al., 2022; Ferber et al., 2024). However, most of the existing works rely on a dark-box model that generates a report without any evidence or thinking process, while many hallucinated content can be find found in the report.

3 Method

Most current chest-X-ray (CXR) report LVLMs rely on global image features, generating descriptions from the whole-image representation alone. This strategy often neglects critical anatomical information. **Clinically, many diseases manifest correlated findings across distinct anatomical structures**, and those structures have systematic, spatially consistent relationships with the visual features in the image. Accurate interpretation of chest radiographs (CXR) often requires a divide-and-conquer, tiered examination of anatomical structures across different levels, rather than treating the image as an undifferentiated whole or focusing solely on a single salient feature. To address the limitations of existing methods, we first construct a Spatial-Temporal anatomical graph whose nodes correspond to anatomical structures 3.1 and enrich it with multi-faceted feature representations; we then interpret this graph using our novel Thinking-With-Graph (TwG) reasoning framework 3.2.

3.1 Spatial-Temporal Anatomical Graph

We define a spatial temporal anatomical graph(STAG) as follows:

$$G = (\mathcal{V}, \mathcal{E}, f) \tag{1}$$

Where $\mathcal{V} = \{v_i^{(t,l)}\}$ is the vertext set and t denotes the temporal index (e.g., examination time) and l denotes the layer in the anatomical hierarchy. $\mathcal{E}$ denote the edge set. The f is the composite feature representation for the nodes. Each node $v_i^{(t,l)} \in \mathcal{V}$ corresponds to a distinct anatomical structure (e.g., lung, heart). Associated with each node is a composite feature $f(v_i^{(t,l)}) = (f_a^{(i)}, f_d^{(i)}, f_{\mathcal{T}}^{(i)})$, the definitions are as followings:

Abnormality Feature $f_a^{(i)} \in \{0, 1\}$ is a binary indicator denoting whether the region is abnormal (1) or normal (0). Following prior work (Tanida et al., 2023; Zhang et al., 2024a), we first employ a detection model to localize anatomical structures of interest in the chest X-ray images, producing bounding boxes, class labels, confidence scores, and associated visual features for each region. Based on these region-level features, a binary classifier is then applied to determine whether each structure is abnormal, upon which abnormal structures are assigned corresponding findings features.

Findings Feature Findings Feature $f_t^{(i)} \in \mathcal{S}$ is a free-text string represents the predicted findings of the region. To keep the generated findings focused on a single region and avoid spill-over, we train a visual-prompt guided model M to generate the structure-level findings. Formally, given an input image x_i and bounding box coordinates b_i, we utilize a visual prompt generator used in VIP (Cai et al., 2024) to obtain various form of visual prompt P_v. These prompts are either circle or rectangle markers that highlight the area defined by the coordinates. Thus, we have the following equation:

$$p_v^{(i)} = \Phi(b_i); p_v^{(i)} \in P_v = P^{\text{rectangle}} \cup P^{\text{ellipse}}$$

where the $P^{\text{rectangle}}$andP^{ellipse} represent the rectangle shape prompt and ellipse shape prompt respectively. After that, the visual prompt $p_v^{(i)}$ will be integrated to the original image x_i and fed into the visual-prompt-aware LVLM $\mathcal{M}$ to generate the finding text $f_{\mathcal{T}}(i)$ for node v_i:

$$f_{\mathcal{T}}(i) = M(p_t, p_v \oplus x_i | \theta_t) \tag{2}$$

Where p_t is the text prompt that we create to instruct the model to generate information only related to the visual-prompted area, the template can be found in the appendix. $p_v \oplus x_i$ indicates the integration of the visual prompt marker and the input image. θ_t represents the parameters of the fine-tuned model $\mathcal{M}$.

Domain Feature Inspired by Chest ImaGenome (Wu et al., 2021), we also adopt the naming scheme from the *RadLex* ontology. Therefore, for each vertex v_i, we have the domain feature: $f_d^{(i)} \in \mathcal{S}$, which is the standardized name and definition of a specific anatomical structure. We use a mapping function c to retrieve the definition and definition z_i from well-established medical ontologies such as UMLS, for the structure related to v_i. To be specific, we have the follows:

$$z_i = c(f_d^{(i)}), z_i \in Z_{umls} \tag{3}$$

Where c indicate the mapping function and Z_{umls} is the ontology. For each node v_i, the mapped description $d^{(i)} = c(f_d^{(i)})$ provides a clinically grounded textual definition of the anatomical entity.

Edge Definitions. The directed edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ consists of three disjoint subsets: We formally define three types of edges in the anatomical graph:

$$\mathcal{E}^H \subseteq \mathcal{V}^{(t,l-1)} \times \mathcal{V}^{(t,l)}, \mathcal{E}^S \subseteq \mathcal{V}^{(t,l)} \times \mathcal{V}^{(t,l)},$$
$$\mathcal{E}^T \subseteq \mathcal{V}^{(t-1,l)} \times \mathcal{V}^{(t,l)}.$$

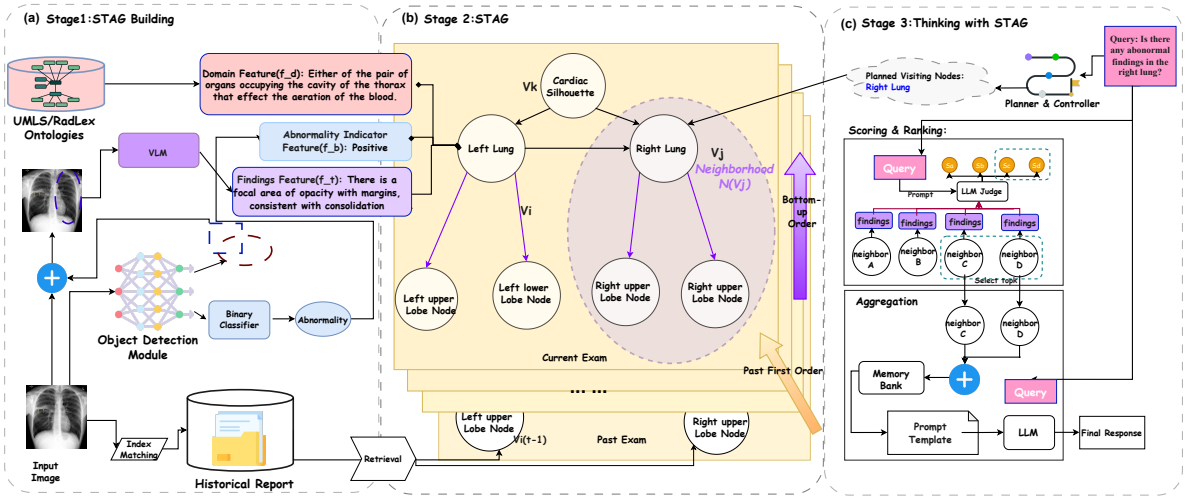


Figure 1: In this figure: we show the framework of building STAG and reasoning with our constructed graph. In Figure 1(a) we show the detail of generate features for a specific node. In Figure 1(b), we have the constructed STAG. Then we show the thinking process with the STAG in (c)

Hierarchy edges ($\mathcal{E}^H$) connect nodes across consecutive hierarchy levels, where each edge $e_{ij}^H = (v_i^{(t,l-1)}, v_j^{(t,l)}) \in \mathcal{E}^H$ indicates that $v_i^{(t,l-1)}$ is the parent of $v_j^{(t,l)}$. This captures part-whole anatomical relationships

Spatial edges ($\mathcal{E}^S$) link nodes within the same hierarchy level, where $e_{ij}^S = \{v_i^{(t,l)}, v_j^{(t,l)}\} \in \mathcal{E}^S$ encodes spatial adjacency or anatomical relatedness (e.g., "left lung" to "right lung").

Temporal edges ($\mathcal{E}^T$) connect the same anatomical region across multiple studies, where $e_{ii}^T = (v_i^{(t-1,l)}, v_i^{(t,l)}) \in \mathcal{E}^T$. These edges only exist when multiple CXR studies from the same patient are available, ensuring longitudinal consistency.

This formulation enables the graph G to encode both compositional and spatial anatomy, facilitating structured reasoning.

3.2 Thinking-with-Graph

To effectively integrate information across multiple spatial and temporal levels within the constructed STAG, thereby enhancing the capacity of LLMs to generate outputs for downstream tasks, it is essential to impose a principled traversal order over its nodes. Such an ordering enables the model to acquire information in a fine-to-coarse manner, reducing the risk of omissions and mitigating premature conclusions. In addition, to help the LLM integrate each node's neighborhood into its reasoning process, we introduce a thinking pattern that serves as the node transformation operation.

Formally, we creatively introduce our **Thinking-with-Graph (TwG) framework**, which uses a controller algorithm 3.2.1 $\mathcal{C}$ to determine a bottom-up execution order $\mathcal{O}_v$ and visiting planner 3.2.3 to define which nodes to expand for a query. To further enhance cross-node reasoning, we incorporate a dedicated **Memory Bank** $\mathcal{MB}$ 3.2.2 into the TwG framework. During traversal, whenever a node is visited, its associated textual feature is stored in $\mathcal{MB}$. As the controller $\mathcal{C}$ expands a node to its neighbors, the framework defines a set of operations 3.2.4 that serve as reasoning primitives, ranking and aggregating information across each node's neighborhood $\mathcal{N}$. The selected neighbor features are then screened and aggregated, and the most relevant representations are appended to $\mathcal{MB}$. This design ensures that information is accumulated sequentially according to the traversal order, enabling both local and global contextual integration. Finally, the complete contents of $\mathcal{MB}$ are consolidated into a structured template, which is passed as a prompt to the LLM for generating task-specific outputs, such as medical report generation or question answering.

3.2.1 Graph Execution Controller

The controller determines how the graph is traversed during reasoning. Intuitively, the traversal is designed to mimic a **DFS-like process**, following a **coarse-to-fine, depth-first schedule**: for each branch, the controller expands nodes down to their fine-grained substructures before backtracking. This ensures that detailed evidence is first gathered at leaf nodes and then propagated upward

to coarser structures, thereby improving the aggregation of information. For example, when expanding the *left lung*, the controller first visits its subregions such as the *left upper lung zone* and *left lower lung zone*, collects their information, and then backtracks to update the representation of the whole *left lung*. To rigorously encode this intuition, we formalize the traversal through an execution order over nodes defined by the orientation of edges in the graph. Formally, we define an execution order $\mathcal{O}v$ over nodes by orienting all edges in the graph. For a *hierarchical edge* $eij^H \in \mathcal{E}^H$ that links a parent node $v_i^{(t,l-1)}$ to a child node $v_j^{(t,l)}$, the child is visited first:

$$\sigma(v_i^{(t,l-1)}) < \sigma(v_j^{(t,l)}) \tag{4}$$

$$\sigma(v_i^{(t,l)}) < \sigma(v_j^{(t-1,l)}) \tag{5}$$

Here, the $\sigma(\cdot)$ function denotes the execution score, with larger values executed earlier.

Equation (4) thus encodes that for hierarchical edges the execution is *depth-first*, i.e., the child node is expanded before the parent node. For a *temporal edge* $e_{ij}^T \in \mathcal{E}^T$, the execution is *old-first*, i.e., the node with the older timestamp is executed before the newer one. Since the *spatial edges* $\mathcal{E}^S$ are undirected, we simply convert them into bidirectional edges for convenience.

3.2.2 Memory Bank

Memory Bank The **Memory Bank** $\mathcal{MB}$ is a dynamic repository designed to accumulate textual evidence during the traversal of the STAG. At each step, when a node v_i is visited, the textual description of v_i together with summarized information from its neighbors is appended into $\mathcal{MB}$: $\mathcal{MB} \leftarrow \mathcal{MB} \cup \{f_{\text{text}}(v_i)\}$. In this way, $\mathcal{MB}$ maintains a sequential record of localized evidence aligned with the execution order $\mathcal{O}_v$. In the final stage, the entire contents of $\mathcal{MB}$ are serialized into a structured prompt, which is then passed to the LLM to generate task-specific outputs:

$$\hat{y} = \text{LLM}(\text{Template}(\mathcal{MB}) \mid \theta_{inf}),$$

where $\hat{y}$ denotes the generated report or answer and θ are the LLM parameters.

3.2.3 Visiting planner.

Unlike conventional traversal methods with fixed termination rules, in our framework the selection criterion is determined dynamically by a large language model (LLM) planner. Given a query q, the planner continually reflects on which nodes of the graph provide the most relevant information and produces a candidate list of nodes to be visited, denoted as $\mathcal{V}_q \subseteq \mathcal{V}$. And, we write: $\mathcal{V}_q = \text{Planner}_\theta(q, G)$ where Planner_θ is an LLM-based function that maps the query q and the underlying STAG G into a prioritized subset of nodes. Traversal is then carried out according to the execution order $\mathcal{O}_v$ defined in the previous section, but restricted to $\mathcal{V}_q$. At the same time, the Memory Bank $\mathcal{MB}$ is continually updated with the textual features of the visited nodes, ensuring that both local and global context are preserved. For comprehensive tasks such as report generation, the planner typically returns $\mathcal{V}_q = \mathcal{V}$, resulting in a complete expansion of the graph. In contrast, for more targeted queries the planner outputs only a task-specific subgraph, e.g.,

$$\mathcal{G}_q = \big(\mathcal{V}_q \cup \mathcal{N}^T(\mathcal{V}_q),\, \mathcal{E}|_{\mathcal{V}_q \cup \mathcal{N}^T(\mathcal{V}_q)}\big),$$

where $\mathcal{N}^T(\mathcal{V}_q)$ denotes the set of temporal descendants of $\mathcal{V}_q$. In this way, the controller executes a bottom-up traversal only over the subgraph $\mathcal{G}_q$, while the Memory Bank $\mathcal{MB}$ accumulates all retrieved textual evidence, ensuring efficiency and retaining the relevant contextual information for answering the query.

3.2.4 Rank and Aggregation Operations

In a graph-based reasoning process, each node represents a unit of information. As the reasoning unfolds, a node needs to integrate evidence from its neighbors and adjust its own state accordingly. To address this, we introduce node-transformation operators, which describe how a node gathers messages from its neighbors and updates its own representation. Common transformation types include:(i) Scoring (T_{score}), which evaluates the plausibility of candidate states;(ii) Ranking (T_{rank}), which orders competing hypotheses;and (iii) Aggregating and Updating (T_{agg}), which merge neighboring features; A node-transformation operator specifies how a node collects information from its neighbors and refines it for reasoning.

Scoring & Ranking In the TwG framework, *scoring* refers to evaluating the quality or relevance of a target node's neighboring nodes $\mathcal{N}(v_i)$ based on their feature representations. This process assigns a score to each neighbor, which reflects its importance for the current reasoning step. *Ranking* then orders these neighbors according to their

Table 1: Direct Generation Results

Method		NLG-Metrics				Clinical Efficacy			
	Explainable	Bleu2	Bleu4	Rouge-L	Meteor	$\text{Prec}_{\text{macro}}$	R_{macro}	$\text{F}_{1,\text{macro}}$	$\text{F}_{1,\text{micro}}$
R2GEN (Chen et al., 2022)	False	0.218	0.103	0.277	0.142	0.331	0.224	0.228	0.346
R2GenCMN	False	0.206	0.094	0.266	0.136	0.334	0.275	0.278	–
KiUT (Huang et al., 2023)	False	0.243	0.113	0.285	0.161	–	–	–	–
COMG (Gu et al., 2023)	False	0.235	0.124	0.283	0.128	0.424	0.291	0.345	–
Damper (Huang et al., 2024)	False	**0.284**	**0.193**	0.301	0.289	0.512	0.473	0.507	–
M2KT*	False	0.240	0.112	0.289	0.147	0.394	0.353	0.372	–
RGRG (Tanida et al., 2023)	True	0.249	0.126	0.264	0.168	0.461	0.475	0.447	0.547
AP-ISG (Zhang et al., 2024a)	True	0.258	0.129	0.264	0.175	0.486	0.493	0.462	0.582
SGRRG (Wang et al., 2025)	True	0.285	0.149	0.326	0.173	0.568	0.513	0.539	–
Ours	True	0.256	0.137	**0.361**	**0.296**	0.582	0.533	**0.557**	**0.625**

Table 2: VQA Results

Method		Mimic-VQA			
	Medical Dedicated	Acc	BLEU-4	Rouge-L	Meteor
InternVL2.5(8B) (Chen et al., 2024b)	No	0.721	0.257	0.610	0.455
Qwen2.5VL(7B) (Bai et al., 2023)	No	0.704	0.256	0.629	0.462
Llava-Med (Li et al., 2024)	Yes	0.770	0.267	0.651	0.473
MedVP (Zhu et al., 2025)	Yes	0.757	0.263	0.637	0.468
Ours	Yes	**0.806**	**0.282**	**0.684**	**0.487**

scores, enabling selective aggregation using only the most informative subset.

Formally, for a target node v_i, we apply a scoring function $\varepsilon_{\text{sc}}(\cdot)$, which can be a fine-tuned large language model (LLM) or other task-specific model, to produce a score for each neighbor:

$$sc_j = \varepsilon_{\text{sc}}(P_{\text{sc}},\ f(v_j) \mid p_\varepsilon), \quad v_j \in \mathcal{N}(v_i), \tag{6}$$

where P_{sc} denotes the scoring prompt or context, $f(v_j)$ is the feature representation of v_j, and p_ε are the parameters of the scoring model.

A ranking function $\gamma(\cdot)$ then sorts the neighbors according to their scores:

$$rk_j = \gamma(\{(v_j, sc_j) \mid v_j \in \mathcal{N}(v_i)\}), \tag{7}$$

where rk_j is the rank position of v_j among all neighbors. Based on this ordering, TwG selects only the top-k neighbors to retain for aggregation:

$$\mathcal{N}_k(v_i) = \{v_j \in \mathcal{N}(v_i) \mid rk_j \leq k\}. \tag{8}$$

This ensures that only the most relevant neighbors contribute to the reasoning process, reducing noise while focusing the LLM on high-value evidence.

Aggregation & Memory Writing After the top-k neighbors $\mathcal{N}_k(v_i)$ are determined, TwG performs aggregation by collecting their textual feature $f_{\mathcal{T}}(v_j)$ and concatenating them into a single contextual sequence. This preserves the original textual form of each neighbor's information and yields a compact summary tailored for the target node. Formally, for a target node v_i, the aggregated representation is defined as:

$$f_{\mathcal{T}}^{\text{agg}}(v_i) = \bigoplus_{v_j \in \mathcal{N}_k(v_i)} f_{\mathcal{T}}(v_j), \tag{9}$$

where $f_{\mathcal{T}}(\cdot)$ denotes the text feature extracted from a node's description, and $\oplus$ is the concatenation operator. Instead of updating the graph topology, the aggregated text is written into the **Memory Bank**: $\mathcal{MB} \leftarrow \mathcal{MB} \cup \{f_t^{\text{agg}}(v_i)\}$.

Through this design, the Memory Bank $\mathcal{MB}$ incrementally accumulates evidence from the most relevant neighborhoods in the traversal order. Later, its contents are serialized into task-specific prompts that condition the LLM for medical report generation or question answering.

4 Experiments

Dataset To validate our framework's performance on different stage, we also utilize the anatomical structure level image-text anotation provided in Chest ImaGenome Dataset (Wu et al., 2021), which is a dataset that derived from the original image-report data from MIMIC-CXR, but only utilizing the frontal view images. The Chest ImaGenome Dataset contains automatically constructed findings and bounding-box annotation for 29 unique anatomical structural in the chest. For fair comparison, we follow RGRG (Tanida et al., 2023) to adopt the official split by the Chest ImaGenome dataset, which has 166,512 training im-

Table 3: Ablation on MRG

Method	NLG-Metrics		Clinical Efficacy	
Modules Used	Rouge-L	Metero	$F_{1,macro}$	$F_{1,micro}$
Direct SFT	0.265	0.196	0.452	0.554
+ SAG (No Temporal Node)	0.338	0.260	0.521	0.580
+ SAG (No Sort & Agg)	0.301	0.231	0.536	0.587
+ Full Model (Ours)	**0.361**	**0.296**	**0.557**	**0.625**

Table 4: Ablation on VQA

Modules Used	Accuracy (%)	Rouge-L
Baseline (w/o any module)	75.2	0.62
+ SAG (No Sort & Agg)	74.7	0.64
+ SAG (No Planner)	75.9	0.61
+ Full Model (Ours)	**80.6**	**0.68**

ages, 23,592 validation images, and 47,389 test images. We also follow the Visual-prompt generation process applied in ViP-LLava (Cai et al., 2024) and MedVP (Zhu et al., 2025), using random shape marker to highlight the anatomical structure area in the CXR images. During inference, we follow the RGRG (Tanida et al., 2023) to employ a pre-trained detector to predict the relevant structure level and generate the visual-prompt accordingly.

For the VQA task, we use the Mimic-Ext-VQA (Bae et al., 2024) to train and test our model. MIMIC-CXR-VQA includes approximately 377K entries derived from the MIMIC-CXR-JPG, MIMIC-IV, and Chest ImaGenome datasets, with question type across:Presence, Anatomy, Attribute, Abnormality, Size, Plane, Gender. And the question forms included Close-Ended (Yes/No, Multi-choice) and Open-Ended questions. This is one of the most comprehensive open-source VQA dataset for CXR modality.

Metrics For medical report generation task, we use the natural language generation(NLG) metrics, such as BLEU (Papineni et al., 2002), Rouge (Lin, 2004) scores to evaluate the final report's quality. And Meteor (Banerjee and Lavie, 2005) is used for structural-level findigns evaluation. Following prior works (Tanida et al., 2023; Huang et al., 2024; Zhang et al., 2024a), we use Clinical Efficacy(CE) to measure the diagnostic accuracy of our generated report. CE metrics evaluate model outputs based on clinically meaningful outcomes, typically by mapping predictions to structured clinical labels and computing precision, recall, and F1 Following prior works (Tanida et al., 2023; Zhang et al., 2024a; Wang et al., 2025), we report both the Micro-4 and Macro F1 values for the CE-metrics

For the VQA task, we do not adopt the multi-label classification formulation employed in the original work of MIMIC-CXR-VQA (Bae et al., 2024), as our primary goal is to evaluate the generative capability of the model. Instead, we follow the commonly used evaluation paradigm in large language models for VQA, where the likelihood of the next-token distribution is used to assess whether the predicted answer matches the ground truth. This setting enables a more faithful reflection of the model's generation ability, rather than restricting it to a fixed classification space.

5 Results Interpretation

Medical Report Generation(MRG) We exclude large foundation models such as MedGemma, and focus our comparison only on methods trained directly on the MIMIC-CXR dataset. For MRG task, we evaluate our method against 6 direct report generation method that lack of explainability and 3 interpretable report generation models that can generate structural level findings. As shown in table 1, our method's peformance is outstanding, either being competitive with or superior in terms of the NLG metrics. For instance, our method achieve State-of-the-Art(SOTA) performance on Rogue-L and Meteor metrics for the report generation task. To be specific, our method achieve SOTA results for Rouge-L and Meteor metrics, resulting a substantial $+10.7\%$ increasing for Rouge-L and nearly $+20\%$ increasing for Meteor metrics. Since ROUGE-L and METEOR are advanced evaluation metrics for natural language generation that move beyond exact n-gram matching (as in BLEU) by incorporating measures of longest common subsequence and semantic similarity, thereby providing a more nuanced assessment of content fidelity and meaning preservation. In Fig **??** six selected anatomical structure respectively and compared with previous work (Tanida et al., 2023; Zhang et al., 2024a) that can perform structural-level findings generation. In TabTab **??**, we also show the generated report and compared with the ground truth as a qualitative study.

As for the CE metrics, also shown in table 1, our model show SOTA performance on both macro and micro F1-scores and leading by a large margin. This superiority on CE metrics is particularly significant, as they directly reflect clinically meaningful correctness, whereas traditional NLG metrics

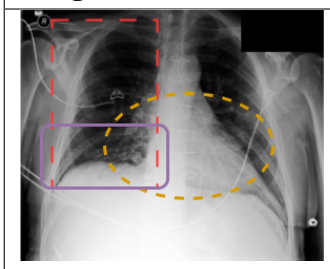
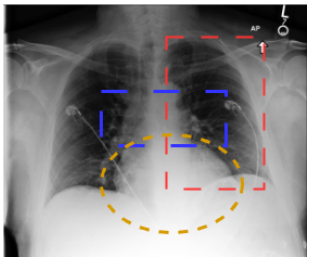
Image	Predicted Report	Ground Truth
	Right basilar opacity has mostly cleared since the prior exam; The cardiac contour is not enlarged; There is no pleural effusion or pneumothorax.	Right basilar opacity has mostly cleared since the prior radiographs. The cardiac, mediastinal and hilar contours are stable. The patient is status post sternotomy. There is no pleural effusion or pneumothorax.
	There is no acute cardiopulmonary abnormality for the cardiac silhouette; mediastinal and hilar contours are unremarkable; Pulmonary vasculature is normal; Lung volumes are slightly low pulmonary vasculature is normal; There is no pleural effusion or pneumothorax no focal consolidation to suggest pneumonia	Lung volumes are slightly low. This accentuates the size of the cardiac silhouette which is likely top normal. Mediastinal and hilar contours are unremarkable.Pulmonary vasculature is normal. No focal consolidation, pleural effusion or pneumothorax is present. There are mild degenerative changes in the thoracic spine. Another ground truth text here.

Table 5: Comparison of predicted reports and ground truth for chest X-ray images.

mainly capture lexical overlap and thus provide a less reliable measure of practical utility.

VQA performance As shown in Table 2, we fine-tuned two general-purpose VLMs and two medical-domain VLMs for comparison with our proposed method. We report the closed-ended accuracy as well as the BLEU-4, METEOR, and ROUGE-L metrics for the open-ended questions. It can be observed that our method achieves a substantial improvement once finer-grained, evidence-based reasoning is introduced into the VQA process. Unlike other models that directly generate answers from the image, our approach selectively filters relevant findings associated with the query and incorporates them into the model's context as part of the chain-of-thought. Although the underlying VLM remains the same, this design yields a notable performance gain.

Ablation Experiments As shown in Tables 3 and 4, we conducted ablation studies on the medical report generation and VQA datasets, respectively. Due to the large scale of the data, our experiments primarily focused on evaluating the performance of the core modules. We first observed that the STAG component plays a crucial role in both report generation and VQA tasks, particularly in integrating information from subordinate nodes. For report generation, if we omit the fusion and selection process and instead concatenate all findings indiscriminately, the NLG metrics drop significantly—likely due to the introduction of redundant and noisy information. Similarly, for the VQA task, the absence of neighbor-node filtering leads to a substantial decrease in accuracy, and in some cases, the performance even falls below that of direct baseline inference, suggesting that excessive irrelevant information can be detrimental. These observations highlight that information fusion and reasoning constitute a critical part of our framework design. In addition, temporal document retrieval and node selection during traversal also play key roles in both tasks, contributing distinctly to their respective performance improvements.

6 Conclusion

In conclusion, our work demonstrates that incorporating explicit graph-based reasoning into medical vision-language models substantially enhances interpretability without sacrificing performance. By unifying spatial-temporal anatomical structure modeling with transparent reasoning through the Thinking-with-Graph paradigm, the proposed framework offers a more controllable and clinically trustworthy foundation for future medical AI.

Limitations and Risk

Our current framework primarily relies on a single imaging modality and an ontology that defines a specific set of anatomical regions. This design choice, while effective in structured chest X-ray analysis, may limit the model's generalizability to other imaging modalities or alternative anatomical taxonomies. Extending the proposed method to more diverse clinical settings will require further adaptation of both the ontology and the reasoning process.

Moreover, during training, we depend on machine-extracted textual findings as the ground truth, which may introduce noise or incomplete supervision. The organ recognition module is also imperfect, and its potential errors could propagate through the reasoning pipeline. At present, our system is intended purely for research and experimental validation; it is not ready for direct clinical deployment. Fully relying on large language models to generate medical reports carries inherent risks, and therefore this work should be viewed as an exploratory study rather than an immediately applicable diagnostic tool.

References

Ammar Ali. 2025. Interpretable deep learning framework for covid-19 detection: Grad-cam integration with pre-trained cnn models on chest x-ray images. *International Journal of Scientific Research in Science, Engineering and Technology*, 12:153–163.

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and 1 others. 2024. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, and 1 others. 2024a. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2022. Generating radiology reports via memory-driven transformer. *Preprint*, arXiv:2010.16056.

Dyke Ferber, Georg Wölflein, Isabella C. Wiest, Marta Ligero, Srividhya Sainath, Narmin Ghaffari Laleh, Omar S. M. El Nahhas, Gustav Müller-Franzes, Dirk Jäger, Daniel Truhn, and Jakob Nikolas Kather. 2024. In-context learning enables multimodal large language models to classify cancer pathology images. *Preprint*, arXiv:2403.07407.

Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. 2023. Complex organ mask guided radiology report generation. *Preprint*, arXiv:2311.02329.

Xiaofei Huang, Wenting Chen, Jie Liu, Qisheng Lu, Xiaoling Luo, and Linlin Shen. 2024. Damper: A dual-stage medical report generation framework with coarse-grained mesh alignment and fine-grained hypergraph matching. *Preprint*, arXiv:2412.14535.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. *Preprint*, arXiv:2306.11345.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *Preprint*, arXiv:1901.07042.

Vanitha Kumar, Mahesh T R, V. Kumar, and Suresh Guluwadi. 2025. Enhanced tuberculosis detection using vision transformers and explainable ai with a grad-cam approach on chest x-rays. *BMC Medical Imaging*, 25.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and 1 others. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. 2023. Medical image understanding with pretrained vision language models: A comprehensive study. *Preprint*, arXiv:2209.15517.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 7433–7442. IEEE.

Jun Wang, Lixing Zhu, Abhir Bhalerao, and Yulan He. 2025. Sgrrg: Leveraging radiology scene graphs for improved and abnormality-aware radiology report generation. *Computerized Medical Imaging and Graphics*, page 102644.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Joy T Wu, Nkechinyere Nneka Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward Christopher Dee, William G Mitchell, Satyananda Kashyap, Andrea Giovannini, and 1 others. 2021. Chest imagenome dataset for clinical reasoning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Yuguang Yang, Tongfei Chen, Haoyu Huang, Linlin Yang, Chunyu Xie, Dawei Leng, Xianbin Cao, and Baochang Zhang. 2025. Prompt as knowledge bank: Boost vision-language model via structural representation for zero-shot medical detection. *Preprint*, arXiv:2502.16223.

Ke Zhang, Yan Yang, Jun Yu, Jianping Fan, Hanliang Jiang, Qingming Huang, and Weidong Han. 2024a. Attribute prototype-guided iterative scene graph for explainable radiology report generation. *IEEE Trans. Med. Imaging*, 43(12):4470–4482.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2024b. Gpt4roi: Instruction tuning large language model on region-of-interest. *Preprint*, arXiv:2307.03601.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Kangyu Zhu, Ziyuan Qin, Huahui Yi, Zekun Jiang, Qicheng Lao, Shaoting Zhang, and Kang Li. 2025. Guiding medical vision-language models with explicit visual prompts: Framework design and comprehensive exploration of prompt variations. *Preprint*, arXiv:2501.02385.

A Appendix

A.1 A. Medical Report Generation with Think-with-Graph

In the scenario of medical report generation, the selection criterion is typically set to *full expansion*, i.e., $\mathcal{V}_q = \mathcal{V}$, so that the entire anatomical graph is traversed. Since the edges of the graph have been oriented into directed form (Eq. 4), the controller executes the traversal in a depth-first, coarse-to-fine manner. At each step of the traversal, the system integrates information from subnodes upward using the transformation operators defined previously.

Specifically, for a leaf node v_i, we first apply the T_{score} and T_{rank} transformations to its neighborhood $\mathcal{N}(v_i)$, retaining the top-k most relevant neighbor nodes and their textual features $f(v)$. For each parent node v_p, we then aggregate over its selected child nodes $\{v_c\}$ using $T_{agg}(\{v_c\})$. In practice, for report generation we find that a simple *concatenation* of the textual findings $f_t^{(j)}$ from child nodes v_c yields superior performance, as the final report can naturally be viewed as a collection of localized findings rather than a compressed vector. Nodes without any detected findings propagate the token *"no anomaly"* upward, ensuring that normal structures are explicitly represented.

During this traversal, each node's final textual description—whether an observed finding or the token *"no anomaly"*—is appended to the **Memory Bank** $\mathcal{MB}$. Thus, $\mathcal{MB}$ accumulates the evidence collected throughout the entire graph in a bottom-up order, preserving both pathological and normal observations across all anatomical regions.

For example, suppose the query requires generating a full chest X-ray report. The controller will first expand fine-grained regions such as the *left upper lung zone* and *left lower lung zone*, collect their textual findings, and then backtrack to update the higher-level node *left lung*. This process continues recursively until reaching the root node (e.g., *thorax*). The concatenated set of textual features $\{f_t^{(i)}\}$ from all nodes is then stored in $\mathcal{MB}$, and finally serialized into a structured prompt:

$$\text{Report} = \text{LLM}\left(\bigoplus_{f_t^{(i)} \in \mathcal{MB}} f_t^{(i)}\right).$$

In this way, the controller-defined traversal, combined with scoring, ranking, aggregation, and the cumulative evidence stored in $\mathcal{MB}$, produces a coherent report that integrates findings from all anatomical regions and conditions the LLM for final report generation.

A.2 Question Answering with Graph

Compared to the MRG task, the VQA task does not require a comprehensive traversal of the entire anatomical graph. Instead, the selection criterion is defined by the LLM-based planner, which identifies a task-specific subgraph sufficient to answer the query. Formally, given a question q, the planner extracts the medical entities mentioned in the query and returns a set of target nodes $\mathcal{V}_q \subseteq \mathcal{V}$:

$$\mathcal{V}_q = \text{Planner}_\theta(q, \mathcal{G}) = \{v_i \in \mathcal{V} \mid f_d(v_i) \cap \mathcal{Q} \neq\},$$

where $\mathcal{Q} = \{q_1, q_2, \ldots, q_k\}$ denotes the set of medical entity terms parsed from q by a domain-specific parser P_{med} (implemented with `spaCy` ()), and $f_d(v_i)$ is the textual descriptor of node v_i.

Once $\mathcal{V}_q$ is identified, the controller initiates a top-down traversal restricted to this subgraph. For each selected node $v_i \in \mathcal{V}_q$, we expand its 1-hop neighborhood $\mathcal{N}^{(1)}(v_i)$, apply T_{score} and T_{rank} to evaluate the relevance of neighbors, and retain the top-k candidates $\mathcal{N}_k(v_i)$. The features of the selected neighbors are then aggregated as:

$$f_t^{\text{agg}}(v_i) = \bigoplus_{v_j \in \mathcal{N}_k(v_i)} f_t(v_j), \tag{10}$$

and appended to the **Memory Bank**:

$$\mathcal{MB} \leftarrow \mathcal{MB} \cup \{f_t^{\text{agg}}(v_i)\}. \tag{11}$$

Unlike the report generation task, here $\mathcal{MB}$ accumulates evidence from multiple scattered subgraphs corresponding to the query-relevant nodes and their neighborhoods. This design allows TwG to preserve localized findings from different anatomical regions without requiring a full traversal of $\mathcal{G}$. Finally, the contents of $\mathcal{MB}$ are serialized and provided to the VQA model together with the question q:

$$\hat{y} = M_{vqa}\left(\bigoplus_{f_t^{\text{agg}}(v_i) \in \mathcal{MB}} f_t^{\text{agg}}(v_i),\ q \mid \theta_{vqa}\right),$$

where $\hat{y}$ denotes the predicted answer and θ_{vqa} represents the parameters of the VQA model M_{vqa} trained separately for this task.

For example, suppose the question is: *"Are there abnormalities in the left hilar structures?"* The

planner will first identify the node corresponding to *left hilar structures* as $\mathcal{V}_q$. The controller then expands its immediate neighborhood $\mathcal{N}^{(1)}(v_{\text{left hilar}})$, applies scoring and ranking transformations, and aggregates the textual features of the most relevant neighbors. Both the description of the target node and the aggregated neighbor information are appended into $\mathcal{MB}$. Since $\mathcal{V}_q$ may contain multiple disjoint nodes, $\mathcal{MB}$ eventually stores a set of scattered evidence blocks, which are collectively passed into M_{vqa} along with the question q to yield the

A.3 Structure Level Results

Region	RL	LL	SP	MED	CS	AB
METEOR	0.214	0.225	0.265	0.281	0.310	0.247

Table 6: Six prominent regions: *right lung* (RL), *left lung* (LL), *spine* (SP), *mediastinum* (MED), *cardiac silhouette* (CS), and *abdomen* (AB).