H^2-MRG: Hierarchical Heterogeneous Graph-Based Medical Report Generation with Fine-Grained Disease-Specific Modeling

Libiao Chen[1] **Dong Nie**[2] **Zhenyu Tang**[1,3,*]
[1]School of Computer Science and Engineering, Beihang University, Beijing, China
[2]TowerCloud Labs
[3]State Key Laboratory of Virtual Reality Technology and Systems,
Beihang University, Beijing, China
[*]tangzhenyu@buaa.edu.cn

Abstract

Vision-language model (VLM)-based medical report generation (MRG) has shown great potential for automating radiology workflows. However, existing VLMs often use coarse fixed-size patches that ignore anatomical boundaries, weakening anatomically grounded visual evidence and structured disease reasoning that are crucial for accurate report generation. In addition, CLIP-style contrastive pretraining treats all unpaired image-report pairs as negatives, introducing false negatives when different patients share similar diseases or findings. To address these issues, we propose H^2-MRG, a hierarchical heterogeneous graph-based MRG framework. Specifically, fine-grained image patches containing complete anatomical structures are extracted. Based on the patches, a hierarchical heterogeneous graph attention network (H^2-GAT) module is presented, where anatomy and disease nodes are connected and aggregate information through anatomy-anatomy (spatial), anatomy-disease (diagnostic), and disease-disease (co-occurrence) relations, enabling anatomically grounded disease reasoning. In addition, an anatomy-level sampling strategy for patch-sentence-level alignment is used to address the false negative problem. Experiments on MIMIC-CXR and IU-Xray datasets demonstrate that our framework outperforms state-of-the-art methods in both natural language generation and clinical efficacy metrics.

1 Introduction

Medical report generation (MRG) aims to produce diagnostic reports from medical images such as chest X-rays (Miura et al., 2021; Wang et al., 2023; Liu et al., 2024; Park et al., 2025). Recent progress in vision-language models (VLMs) has improved report fluency and factuality (Hou et al., 2025; Park et al., 2025; Chen et al., 2025; Xiao et al., 2025; Salhi and Akhloufi, 2026; Liu et al., 2026; Chen et al., 2026). As a result, MRG is increasingly recognized as a testbed for clinically grounded multimodal reasoning, where the central challenge is not merely to produce fluent text, but to ensure that each clinical statement is grounded in anatomically localized visual evidence and supported by structured disease reasoning.

Patch-based ViT encoders partition images into uniform fixed-size tokens (Dosovitskiy et al., 2021); by construction, these tokens do not explicitly preserve anatomical boundaries. As shown in Fig. 1 top, this uniform partitioning can fragment anatomical structures such as the right lung across multiple patches, each capturing only a partial view of the target structure. This weakens the grounding of clinical statements in anatomically localized visual evidence.

Recent studies have therefore explored fine-grained image patches that correspond to meaningful anatomical structures (e.g., left lung, right clavicle) to address this limitation (Tanida et al., 2023; Dalla Serra et al., 2023; Chen et al., 2024). However, preserving complete anatomical boundaries is only the first step toward clinically reliable generation. The critical remaining challenge is to transform anatomy-level visual evidence into disease-level evidence, which requires integrating spatial relations among anatomical structures, diagnostic relations between anatomy and diseases, and co-occurrence relations among diseases. As illustrated in Fig. 1 middle, existing methods typically derive disease cues through simple feature fusion or shallow supervision without explicitly modeling structured disease reasoning (Tanida et al., 2023; Dalla Serra et al., 2023; Li et al., 2023b; Chen et al., 2024), leaving a structural gap between anatomy-aware perception and the structured disease reasoning that clinically reliable report generation demands.

Standard CLIP-style contrastive pretraining treats non-paired samples in a batch as nega-

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 51002–51015
October 24-29, 2026 ©2026 Association for Computational Linguistics

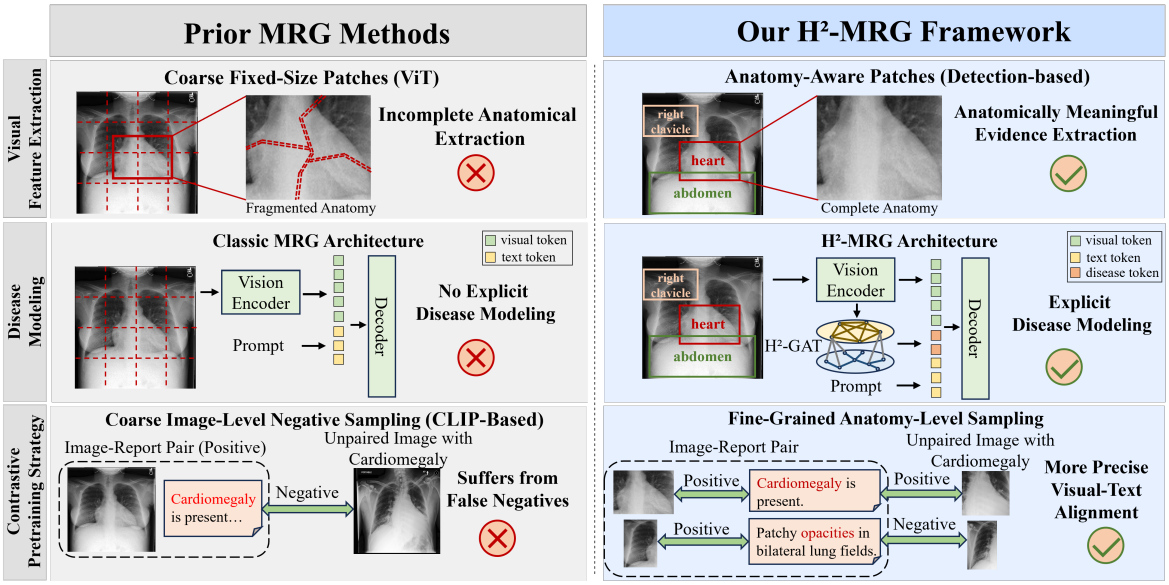


Figure 1: Comparison between prior MRG methods and the proposed H^2-MRG framework in three aspects. Top: prior methods rely on coarse and fixed-size patches, leading to incomplete anatomical evidence extraction, whereas our detection-based anatomy-aware patches preserve complete anatomical structures. Middle: classic MRG architectures lack structured disease reasoning, while H^2-MRG bridges this gap with H^2-GAT. Bottom: coarse image-level negative sampling in CLIP-style contrastive pretraining causes false negatives, while our framework adopts fine-grained anatomy-level sampling.

tives (Radford et al., 2021). In medical imaging, this can create false negatives when different patients share findings (Wang et al., 2022c; Ko and Park, 2025). As illustrated in Fig. 1 bottom, an image exhibiting cardiomegaly may be treated as a negative for a report describing the same finding, introducing false-negative supervision that can degrade visual-text alignment.

To address these challenges, we propose H^2-MRG, a hierarchical heterogeneous graph-based medical report generation framework. Instead of using fixed-size image patches, we adopt an anatomy detection module to extract fine-grained patches that preserve complete anatomical structures and provide anatomy-aware visual tokens. Based on these patches, a hierarchical heterogeneous graph attention network (H^2-GAT) module is presented, where anatomy and disease nodes interact through spatial, diagnostic, and co-occurrence relations to explicitly model the structured reasoning pathway from anatomical evidence to disease inference. In addition, to mitigate false negatives in CLIP-style contrastive pretraining, we introduce an anatomy-level sampling strategy that encourages precise visual-text alignment at the patch-sentence level rather than the image-report level. Experiments demonstrate that H^2-MRG outperforms state-of-the-art methods in both natural language generation and clinical efficacy metrics on MIMIC-CXR and IU-Xray.

The main contributions of our work are listed as follows:

- We introduce anatomy-aware visual tokenization for MRG that extracts fine-grained image patches preserving complete anatomical structures, replacing coarse and fixed-size patches with anatomy-aware visual tokens.
- We propose H^2-GAT, a hierarchical heterogeneous graph attention network that explicitly models the structured reasoning pathway from anatomical evidence to disease inference through spatial, diagnostic, and co-occurrence relations.
- We introduce an anatomy-level sampling strategy for CLIP-style contrastive pretraining to address the false negative problem in image-level negative sampling.
- Extensive experiments on MIMIC-CXR and IU-Xray demonstrate state-of-the-art performance in both natural language generation and clinical efficacy, with ablation studies confirming the contribution of each component.

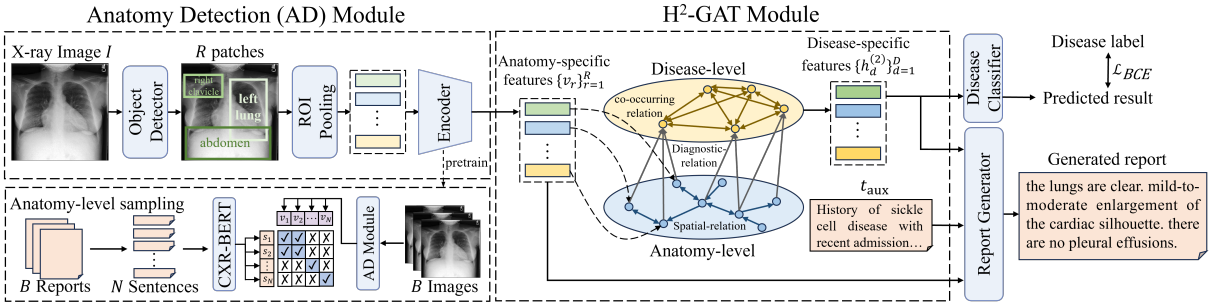


Figure 2: Overview of H^2-MRG. The Anatomy Detection (AD) module extracts fine-grained anatomical patches and encodes them into anatomy-specific features $\{v_r\}_{r=1}^{R}$ via a pretrained ViT encoder. The H^2-GAT module connects anatomy and disease nodes through spatial, diagnostic, and co-occurrence relations to derive disease-specific features $\{h_d^{(2)}\}_{d=1}^{D}$, supervised by a multi-label disease classifier. The Report Generator produces medical reports conditioned on $\{v_r\}_{r=1}^{R}$, $\{h_d^{(2)}\}_{d=1}^{D}$, and auxiliary diagnostic information (e.g., indication and reason for exam).

2 Related Work

2.1 Medical Report Generation

MRG systems generally consist of a visual encoder and a text decoder. Early work improved report accuracy through cross-modal memory networks (Chen et al., 2021), region-level alignment (You et al., 2021; Wang et al., 2022b), and knowledge-guided generation (Yang et al., 2023, 2022). Subsequent methods strengthened structured disease reasoning via observation-structure modeling (Hou et al., 2023b,a) and disease-driven prompting (Jin et al., 2024). More recent systems further improve factual grounding through large language models, retrieval augmentation, and stronger structured disease reasoning (Liu et al., 2024; Hou et al., 2024; Zhou and Wang, 2024; Hou et al., 2025; Park et al., 2025; Chen et al., 2025; Xiao et al., 2025; Wu et al., 2026; Liu et al., 2026; Chen et al., 2026). Despite this progress, these methods share a common visual bottleneck: the encoder partitions images into coarse fixed-size patches that ignore anatomical boundaries, limiting the extraction of meaningful anatomy- and disease-specific features.

2.2 Semantic-Aware Visual Processing in MRG

To address the coarse-patch limitation, recent studies explore fine-grained anatomical patches for report generation. AlignTransformer (You et al., 2021) aligns anatomical regions with disease tags, Tanida et al. (Tanida et al., 2023) detect meaningful structures and generate anatomy-specific sentences, and Müller et al. (Müller et al., 2022; Müller et al., 2024) learn localized image-report representations for fine-grained grounding. Li et al. (Li et al., 2023b) jointly optimize global and local alignment, and Chen et al. (Chen et al., 2024) propose adaptive alignment without predefined structures. Despite these advances, existing methods typically extract disease-specific features through simple feature fusion or shallow supervision, without explicitly modeling structured disease reasoning from anatomical evidence to disease inference (You et al., 2021; Tanida et al., 2023; Dalla Serra et al., 2023; Li et al., 2023b; Chen et al., 2024).

2.3 Contrastive Learning in Pretraining

CLIP-style contrastive pretraining has been widely adopted in MRG to learn aligned visual-text representations (Radford et al., 2021; Zhang et al., 2022; Li et al., 2023a; Park et al., 2025; Liu et al., 2025), with global-local strategies such as GLoRIA and MGCA providing multi-granularity alignment (Huang et al., 2021; Wang et al., 2022a). To alleviate false negatives, MedCLIP (Wang et al., 2022c) introduces semantic matching based on medical entity similarity, and Ko et al. (Ko and Park, 2025) incorporate dynamic soft labels based on multi-source similarities. However, these approaches remain constrained to image-report-level alignment and cannot fundamentally eliminate false negatives arising from inter-patient disease overlap.

3 Method

The overview of the proposed H^2-MRG is illustrated in Fig. 2. Specifically, an anatomy detection (AD) module first extracts fine-grained image patches that preserve complete anatomical structures, which are then encoded into anatomy-specific features by a ViT encoder. The anatomy-specific features are passed to a hierarchical hetero-

geneous graph attention network (H^2-GAT), where anatomy and disease nodes interact through spatial, diagnostic, and co-occurrence relations to derive disease-specific features. A disease classifier further supervises H^2-GAT to learn discriminative representations. Based on the anatomy-specific features from the AD module, the disease-specific features from H^2-GAT, and auxiliary diagnostic information, the final reports are generated by a report generator. To pretrain the ViT encoder, an anatomy-level sampling strategy defines positive and negative pairs at the patch-sentence level rather than the image-report level, mitigating false negatives in CLIP-style contrastive pretraining. Details are presented in the following subsections.

3.1 Fine-grained Image Patching and Pretraining of Image Encoder

In our framework, for each input medical image (e.g., chest X-ray image I), a Faster R-CNN (Ren et al., 2015)-based anatomy detection module is used to locate semantically complete anatomical structures. Following the anatomical ontology defined in Chest ImaGenome (Wu et al., 2021), $R = 29$ distinct anatomical structures (e.g., left lung, cardiac silhouette) are determined with corresponding bounding boxes (i.e., fine-grained image patches). For each image patch of a distinct anatomical structure $r \in \{1, \ldots, R\}$, initial anatomy-specific visual features v'_r can be extracted:

$$v'_r = \begin{cases} \text{R_P}(r), & \text{if patch } r \text{ is detected,} \\ v_{\text{pad}}, & \text{otherwise,} \end{cases} \tag{1}$$

where R_P$(\cdot)$ denotes the ROI pooling operation of the detected image patch r, and v_{pad} is a learned embedding of a missing anatomical structure. The sequence $\mathbf{V}_{\text{ini}} = [v'_1; \ldots; v'_R] \in \mathbb{R}^{R \times C_v}$ (where C_v is the feature dimension), augmented with learnable region-identity embeddings, is fed into a ViT encoder (Dosovitskiy et al., 2021) to yield the final anatomy-specific visual features $\mathbf{V} = [v_1; \ldots; v_R] \in \mathbb{R}^{R \times C_v}$ with contextualized information.

We pretrain the ViT encoder via CLIP-style contrastive pretraining with an anatomy-level sampling strategy for better visual-text alignment. Rather than treating all unpaired images as negatives, we parse each report into anatomy-specific sentences (Wu et al., 2021) and align each sentence with its corresponding patch, forming patch-sentence pairs (v_i, s_i) where v_i is the anatomy-specific feature extracted by the ViT encoder and s_i is the textual feature encoded by a pretrained medical BERT encoder (Boecking et al., 2022). Each pair is assigned an anatomy ID $a_i \in \{1, \ldots, R\}$ (e.g., heart = 1 and left lung = 2) and anatomy state $\sigma_i \in \{\text{normal}, \text{abnormal}\}$ via rule-based extraction (Wu et al., 2021). For each anchor pair (v_i, s_i) in a batch, the positive and negative index sets are defined as:

$$\mathbf{P}_i = \{j \mid (a_j = a_i) \wedge (\sigma_j = \sigma_i)\}, \tag{2}$$

$$\mathbf{N}_i = \{j \mid (a_j \neq a_i) \vee (\sigma_j \neq \sigma_i)\}. \tag{3}$$

Therefore, pairs sharing the same anatomy ID and anatomy state are treated as positives, whereas the others are treated as negatives. This design allows cross-sample positive matching beyond the paired sample from the same study.

The contrastive loss is computed using symmetric multi-positive InfoNCE (Radford et al., 2021):

$$\mathcal{L}_{\text{CL}} = \frac{1}{2}\left(\mathcal{L}_{\text{v2s}} + \mathcal{L}_{\text{s2v}}\right), \tag{4}$$

For the visual-to-text direction, each anatomy-specific visual feature v_i is used as the anchor and all sentence features $\{s_j\}_{j \in \mathbf{P}_i}$ are treated as positives. Let $s_{ij} = \cos(v_i, s_j)/\tau$ denote the temperature-scaled similarity:

$$\mathcal{L}_{\text{v2s}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\sum_{j \in \mathbf{P}_i} \exp(s_{ij})}{\sum_{j \in \mathbf{P}_i \cup \mathbf{N}_i} \exp(s_{ij})}, \tag{5}$$

where N is the number of patch-sentence pairs in a batch, $\tau = 0.07$ is the temperature, and $\cos(\cdot, \cdot)$ denotes cosine similarity. $\mathcal{L}_{\text{s2v}}$ is defined symmetrically with s_i as the anchor over visual features.

3.2 Hierarchical Heterogeneous Graph Attention Network (H^2-GAT)

Anatomy-specific features capture localized visual evidence, but generating clinically reliable reports requires structured disease reasoning to produce disease-level conclusions. To this end, we propose the H^2-GAT module to derive effective disease-specific features from the anatomy-specific features produced by the ViT encoder. The key motivation for a graph-based design is that anatomy-disease relations constitute clinically established priors stable across patients. Encoding these priors as graph topology provides a structural inductive bias, allowing the model to learn attention over a fixed relational scaffold rather than re-discovering

the reasoning structure from limited data. H^2-GAT constructs a hierarchical heterogeneous graph (H^2 graph) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with two types of nodes: anatomy nodes and disease nodes. Inspired by the clinical diagnostic workflow, in which radiologists follow a systematic bottom-up analysis from spatial contextualization of anatomy to diagnostic reasoning and finally to co-occurrence-aware refinement among diseases, the H^2 graph is composed of two layers: the low anatomy layer and the high disease layer, where nodes are connected and aggregate information through spatial, diagnostic, and co-occurrence relations.

The low anatomy layer contains R anatomy nodes $\mathcal{V}_A = \{v_r^{(0)}\}_{r=1}^{R} \subset \mathcal{V}$, each of which represents an anatomical structure and is initialized by the corresponding anatomy-specific feature from the ViT encoder: $v_r^{(0)} = v_r$, $r = 1, \ldots, R$. A critical observation in clinical practice is that pathological changes rarely occur in isolation, as diseases often induce spatial effects on adjacent anatomical structures. For instance, cardiomegaly (enlarged heart) may compress surrounding mediastinal structures and affect adjacent lung regions. Therefore, anatomy nodes are connected by edges $\mathcal{E}_{AA} \subset \mathcal{E}$ through the spatial relation, where each node aggregates information from its K-nearest spatial neighbors ($K = 5$ in our study). Spatial distances are computed between centers of corresponding image patches, and an adjacency matrix $\mathbf{A}_{AA} \in \{0,1\}^{R\times R}$ encodes connections, with $\mathbf{A}_{AA}(i,j) = 1$ indicating spatial adjacency between anatomical structures i and j, and 0 otherwise.

The high disease layer comprises D disease nodes $\mathcal{V}_D = \{h_d^{(0)}\}_{d=1}^{D} \subset \mathcal{V}$, each representing a disease category from CheXpert (Irvin et al., 2019) (e.g., pneumonia, cardiomegaly) and initialized with a learnable embedding $h_d^{(0)} \in \mathbb{R}^{C_v}$. A key clinical insight is that diseases often co-occur due to shared etiologies or pathophysiological cascades. For example, congestive heart failure frequently coexists with pulmonary edema and pleural effusion due to fluid redistribution. Therefore, disease nodes are connected by edges $\mathcal{E}_{DD} \subset \mathcal{E}$ through the co-occurrence relation. Specifically, based on disease co-occurrence derived from CheXpert labels in training-split reports, we use the Jaccard similarity coefficient (Jaccard, 1901) to quantify the co-occurrence relation between each pair of diseases, which is defined as:

$$w_{ij} = \frac{|\{\text{Reports}|\text{having } d_i \wedge d_j\}|}{|\{\text{Reports}|\text{having } d_i \vee d_j\}|}, \tag{6}$$

where the numerator and denominator represent the number of reports contained in each set, d_i and d_j stand for different diseases. The results are represented by a weighted adjacency matrix $\mathbf{A}_{DD} \in [0,1]^{D\times D}$, where $\mathbf{A}_{DD}(i,j)$ represents the strength of the co-occurrence relation between diseases i and j.

In addition, anatomy-disease edges $\mathcal{E}_{AD} \subset \mathcal{E}$ bridge the two layers by encoding the diagnostic knowledge that each disease manifests in specific anatomical localizations. For example, pneumonia primarily affects lung parenchyma regions, and cardiomegaly is evidenced by cardiac silhouette enlargement. By modeling such anatomical localization of diseases, the graph enables evidence-based disease inference from relevant anatomical observations while suppressing irrelevant features. Following previous work (Wu et al., 2021), we extract corresponding anatomical structures for each disease from reports in the training split to construct an adjacency matrix $\mathbf{A}_{AD} \in \{0,1\}^{R\times D}$, where $\mathbf{A}_{AD}(r,d) = 1$ indicates that anatomical structure r is relevant to disease d, and 0 otherwise.

Based on the H^2 graph constructed from clinically established and relatively stable anatomy and disease relations, graph attention operations (GAT) (Velickovic et al., 2018) are employed to progressively aggregate information from anatomy nodes to disease nodes. Due to the heterogeneous nature of the H^2 graph, where nodes represent different semantic entities (anatomy and disease) and edges encode distinct relational semantics (spatial, diagnostic, and co-occurrence), each of the three relation types is processed with dedicated attention parameters, enabling structured disease reasoning from anatomical evidence to disease-level inference.

The information aggregation based on the three types of relations consists of three corresponding steps, as illustrated in Fig. 3. In the first step, anatomy-specific features of anatomy nodes are enhanced with spatial contextualization through the spatial relation. For each anatomy node r, information is aggregated from its spatially adjacent neighbors, which is defined as:

$$v_r^{(1)} = \sigma\left(\sum_{r'\in\mathcal{N}(r)} \alpha_{rr'}^{AA} \mathbf{W}_{AA} v_{r'}^{(0)}\right), \tag{7}$$

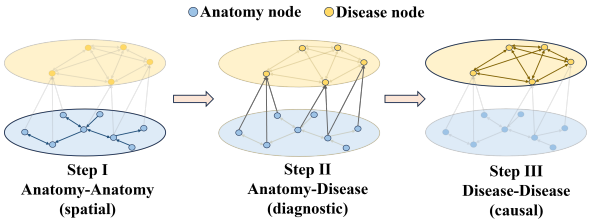


Figure 3: Information aggregation in H^2-GAT across three steps corresponding to the spatial, diagnostic, and co-occurrence relations.

where $\mathcal{N}(r)$ denotes the set of spatial neighbors of anatomy node r (defined by $\mathbf{A}_{AA}$), $\mathbf{W}_{AA} \in \mathbb{R}^{C_v \times C_v}$ is a learnable transformation matrix specific to spatial relations, and σ is a non-linear activation function. The attention coefficient $\alpha_{rr'}^{AA}$ is computed by:

$$\begin{aligned} \alpha_{rr'}^{AA} &= \frac{\exp(e_{rr'}^{AA})}{\sum_{r'' \in \mathcal{N}(r)} \exp(e_{rr''}^{AA})}, \\ e_{rr'}^{AA} &= \text{LeakyReLU}\Big(\mathbf{a}_{AA}^{\top}[\mathbf{W}_{AA} v_r^{(0)} \\ &\qquad \| \, \mathbf{W}_{AA} v_{r'}^{(0)}]\Big), \end{aligned} \tag{8}$$

where $\mathbf{a}_{AA} \in \mathbb{R}^{2C_v}$ is a learnable attention vector for spatial relations and $\|$ denotes concatenation. This stage produces spatially contextualized anatomy features $\{v_r^{(1)}\}_{r=1}^{R}$.

In the second step, anatomy features are transferred to disease nodes through the diagnostic relation. Since features in anatomy and disease nodes reside in different feature spaces, a dual-projection mechanism is employed to project both types of features into a common latent space before computing attention. For each disease node d, we aggregate features from relevant anatomy nodes, which is defined as:

$$h_d^{(1)} = \sigma\left(\sum_{r:\, \mathbf{A}_{AD}(r,d)=1} \alpha_{rd}^{AD} \mathbf{W}_{AD,A} v_r^{(1)}\right), \tag{9}$$

where the attention coefficient is computed as:

$$\begin{aligned} \alpha_{rd}^{AD} &= \frac{\exp(e_{rd}^{AD})}{\sum_{r':\mathbf{A}_{AD}(r',d)=1} \exp(e_{r'd}^{AD})}, \\ e_{rd}^{AD} &= \text{LeakyReLU}\Big(\mathbf{a}_{AD}^{\top}[\mathbf{W}_{AD,A} v_r^{(1)} \\ &\qquad \| \, \mathbf{W}_{AD,D} h_d^{(0)}]\Big), \end{aligned} \tag{10}$$

where $\mathbf{W}_{AD,A} \in \mathbb{R}^{C_v \times d_h}$ projects anatomy features and $\mathbf{W}_{AD,D} \in \mathbb{R}^{C_v \times d_h}$ projects disease embeddings into a common latent space of dimension d_h, and $\mathbf{a}_{AD} \in \mathbb{R}^{2d_h}$ is a learnable attention vector specific to diagnostic relations. The output consists of evidence-based disease features $\{h_d^{(1)}\}_{d=1}^{D}$.

In the third step, the disease-specific features obtained in the second step are further refined among diseases to capture their co-occurrence patterns through the co-occurrence relation. Specifically, for each disease node d, we aggregate information as defined by:

$$h_d^{(2)} = \sigma\left(\sum_{d'=1}^{D} \alpha_{dd'}^{DD} \mathbf{W}_{DD} h_{d'}^{(1)}\right), \tag{11}$$

where the attention coefficient incorporates the precomputed co-occurrence weights:

$$\begin{aligned} \alpha_{dd'}^{DD} &= \frac{\exp(e_{dd'}^{DD})}{\sum_{d''=1}^{D} \exp(e_{dd''}^{DD})}, \\ e_{dd'}^{DD} &= w_{dd'} \, \text{LeakyReLU}\Big(\mathbf{a}_{DD}^{\top}[\mathbf{W}_{DD} h_d^{(1)} \\ &\qquad \| \, \mathbf{W}_{DD} h_{d'}^{(1)}]\Big), \end{aligned} \tag{12}$$

where $w_{dd'}$ is the Jaccard co-occurrence weight, $\mathbf{W}_{DD} \in \mathbb{R}^{C_v \times C_v}$ is a learnable transformation matrix for co-occurrence relations, and $\mathbf{a}_{DD} \in \mathbb{R}^{2C_v}$ is a learnable attention vector. This produces the final disease-specific features $\{h_d^{(2)}\}_{d=1}^{D}$, grounded in anatomical evidence and refined by inter-disease co-occurrence.

To further guide H^2-GAT toward discriminative representations, a disease classifier is added to perform multi-label disease classification based on the resulting disease-specific features $\{h_d^{(2)}\}_{d=1}^{D}$. Specifically, for each disease d, a corresponding classifier Φ_d is used to predict the presence of d:

$$\hat{y}_d = \Phi_d(h_d^{(2)}) \in [0, 1], \tag{13}$$

where $\hat{y}_d$ is the predicted probability. The multi-label classification loss is computed using binary cross-entropy:

$$\begin{aligned} \mathcal{L}_{\text{BCE}} = -\sum_{d=1}^{D} \big[& y_d \log(\hat{y}_d) \\ & + (1 - y_d) \log(1 - \hat{y}_d) \big], \end{aligned} \tag{14}$$

where $y = [y_1, y_2, \ldots, y_D] \in \{0, 1\}^D$ represents the ground truth disease labels.

3.3 Report Generation

The final medical report is generated autoregressively by a Transformer decoder (Wang et al.,

2023; Liu et al., 2024) conditioned on the anatomy-specific visual features $\{v_r\}_{r=1}^{R}$ from the ViT encoder, the disease-specific features $\{h_d^{(2)}\}_{d=1}^{D}$ from H^2-GAT, and the auxiliary diagnostic information t_{aux} provided by the original dataset. The decoder is trained with a next-token prediction objective:

$$\mathcal{L}_{\text{CE}} = -\sum_{t=1}^{T} \log p\big(y_t \mid y_{<t}, \{v_r\}_{r=1}^{R}, \{h_d^{(2)}\}_{d=1}^{D}, t_{\text{aux}}\big), \tag{15}$$

where $\{y_t\}_{t=1}^{T}$ denotes the ground-truth report tokens.

4 Experiments

We evaluate H^2-MRG on MIMIC-CXR (Johnson et al., 2019) and IU-Xray (Demner-Fushman et al., 2016). Specifically, MIMIC-CXR contains chest X-ray images paired with radiology reports and auxiliary diagnostic information (e.g., indication and reason for exam), with anatomy detection leveraging the Chest ImaGenome dataset (Wu et al., 2021) bounding boxes for 29 anatomical structures. Following the split used in prior work (Jin et al., 2024), we use AP/PA view images with 222,322 images for training, 1,855 for validation, and 3,034 for testing. The graph priors $\mathbf{A}_{AD}$ and $\mathbf{A}_{DD}$ are constructed only from reports in the MIMIC-CXR training split and then kept fixed during validation and testing. For IU-Xray, we extract AP/PA view images in the same way, yielding 2,086 images that are used as an external independent test set.

The results of H^2-MRG are evaluated using both natural language generation (NLG) and clinical efficacy (CE) metrics. NLG metrics (BLEU-n, METEOR, ROUGE-L) assess linguistic quality, and CE metrics measure clinical accuracy via the CheXpert labeler (Irvin et al., 2019) on 14 diseases. Our CE scores are micro-averaged; baseline values follow their cited papers' original protocols.

4.1 Implementation Details

Our H^2-MRG follows a three-stage training pipeline. A Faster R-CNN anatomy detector is first trained to localize 29 anatomical regions. The ViT encoder is then pretrained with the proposed anatomy-level sampling strategy under CLIP-style contrastive pretraining. Finally, the full pipeline is jointly fine-tuned with cross-entropy and multi-label BCE losses. Implementation details are provided in Appendix A.

4.2 Evaluation Results of Report Generation

As shown in Table 1, we compare H^2-MRG with six state-of-the-art MRG methods: RGRG (Tanida et al., 2023), CMCRL (Chen et al., 2025), B-LLM (Liu et al., 2024), MLRG (Liu et al., 2025), ICON (Hou et al., 2024) and DART (Park et al., 2025). H^2-MRG achieves the best reported results on most NLG metrics across both datasets and the highest reported clinical-efficacy F1 on MIMIC-CXR. On IU-Xray, it also shows strong transfer performance, indicating that anatomically grounded perception and structured disease reasoning improve not only report fluency but also clinically relevant generalization. A node-level attention visualization is provided in Appendix D.

4.3 Ablation Study

Table 2 reports component-level ablation results on MIMIC-CXR. Model A is the baseline composed of a standard ViT encoder and Transformer decoder. Model B adds anatomy-level sampling (ALS) for patch-sentence-level contrastive learning and brings substantial improvements across both NLG and CE metrics (F1: $0.435 \rightarrow 0.477$), validat-

Table 1: Evaluation results on MIMIC-CXR and IU-Xray. Baseline values follow their original protocols; unavailable metrics are "–". Best and second-best values are in **bold** and underlined.

Dataset	Model	NLG metrics ↑						CE metrics ↑		
		B-1	B-2	B-3	B-4	MTR	R-L	P	R	F1
MIMIC-CXR	RGRG (Tanida et al., 2023)	0.373	0.249	0.175	0.126	0.168	0.264	0.461	0.475	0.447
	CMCRL (Chen et al., 2025)	0.400	0.245	0.165	0.119	0.150	0.280	0.489	0.340	0.401
	B-LLM (Liu et al., 2024)	0.402	0.262	0.180	0.128	0.175	0.291	0.465	0.482	0.473
	MLRG (Liu et al., 2025)	0.411	0.277	0.204	0.158	0.176	0.320	0.549	0.468	0.505
	ICON (Hou et al., 2024)	0.429	0.266	0.178	0.126	0.170	0.287	0.445	0.505	0.464
	DART (Park et al., 2025)	**0.437**	0.279	0.191	0.137	0.175	0.310	0.520	**0.546**	0.533
	Ours	0.434	**0.295**	**0.208**	**0.161**	**0.185**	**0.338**	**0.556**	0.543	**0.549**
IU-Xray	CMCRL (Chen et al., 2025)	0.505	0.334	0.245	0.190	0.210	0.394	-	-	-
	B-LLM (Liu et al., 2024)	0.499	0.323	0.238	0.184	0.208	0.390	-	-	-
	ICON (Hou et al., 2024)	–	–	–	0.098	–	0.320	–	–	–
	DART (Park et al., 2025)	0.486	**0.348**	0.265	0.208	0.205	**0.411**	-	-	-
	Ours	**0.511**	0.341	**0.268**	**0.210**	**0.213**	0.398	-	-	-

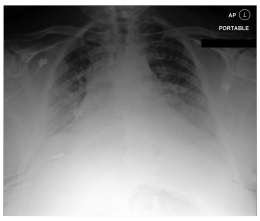
Original Image	
	**Ground-Truth:** No evidence of free air is seen beneath the diaphragms. The cardiomediastinal silhouette remains enlarged. Lung volumes are relatively low and there are increased interstitial markings bilaterally suggesting mild-to-moderate vascular congestion/edema. Blunting of the costophrenic angles is seen which may relate to overlying soft tissue although small pleural effusions may be present.
	Ours w/o H²-GAT: Evaluation is quite limited due to low lung volumes and underpenetrated technique. The heart appears enlarged though this could be in part due to technique. There is no definite sign of free air below the right hemidiaphragm. No large pneumothorax or pneumomediastinum is seen.
	Ours: There are moderate bilateral pleural effusions with overlying atelectasis underlying consolidation is not excluded. The cardiac silhouette is enlarged. There is prominence and indistinctness of the hila suggesting pulmonary vascular engorgement. No pneumothorax is seen.

Figure 4: Case study comparing the baseline (Ours w/o H^2-GAT) and H^2-MRG. Green marks findings consistent with the ground truth; red marks incorrect findings.

Table 2: Evaluation results of ablation study. ALS: anatomy-level contrastive sampling for encoder pretraining; MDC: multi-label disease classifier.

Model	Components			NLG metrics ↑						CE metrics ↑		
	ALS	H^2-GAT	MDC	B-1	B-2	B-3	B-4	MTR	R-L	P	R	F1
A	✗	✗	✗	0.387	0.253	0.168	0.117	0.143	0.283	0.448	0.423	0.435
B	✓	✗	✗	0.408	0.272	0.187	0.136	0.157	0.312	0.488	0.467	0.477
C	✓	✓	✗	0.423	0.286	0.199	0.146	0.171	0.332	0.528	0.512	0.520
Ours	✓	✓	✓	0.434	0.295	0.208	0.161	0.185	0.338	0.556	0.543	0.549

ing that the anatomy-level sampling strategy establishes precise visual-text alignment. Model C further incorporates the H^2-GAT and yields significant CE gains (F1: $0.477 \rightarrow 0.520$), demonstrating the effectiveness of structured disease reasoning. The complete H^2-MRG additionally includes a multi-label disease classifier (MDC), further improving F1 from 0.520 to 0.549, showing that multi-label classification supervision enhances disease-specific representations for accurate report generation.

Table 3 reports relation-level ablation on MIMIC-CXR, where each edge type ($\mathcal{E}_{AA}$, $\mathcal{E}_{DD}$, $\mathcal{E}_{AD}$) is removed from H^2-GAT. Removing $\mathcal{E}_{AA}$ or $\mathcal{E}_{DD}$ leads to moderate degradation, suggesting both spatial contextualization and co-occurrence-based refinement are beneficial. Removing $\mathcal{E}_{AD}$ (i.e., removing H^2-GAT entirely) causes the largest drop, indicating that direct anatomy-to-disease evidence transfer is essential.

Table 3: Relation-level ablation results.

Configuration	B-4	P	R	F1
H^2-MRG	**0.161**	**0.556**	**0.543**	**0.549**
w/o $\mathcal{E}_{AA}$	0.155	0.542	0.528	0.535
w/o $\mathcal{E}_{DD}$	0.157	0.548	0.535	0.541
w/o $\mathcal{E}_{AD}$	0.136	0.488	0.467	0.477

4.4 Case Study

Fig. 4 compares reports generated by H^2-MRG and the baseline without H^2-GAT. The chest X-ray image shows blunting of costophrenic angles suggestive of pleural effusions with vascular congestion. H^2-MRG correctly identifies "moderate bilateral pleural effusions" and "pulmonary vascular engorgement," consistent with the ground truth findings, while the baseline without H^2-GAT misses both critical findings. This example demonstrates that H^2-GAT enables structured disease reasoning by capturing co-occurrence patterns through its hierarchical heterogeneous graph structure.

5 Conclusion

We proposed H^2-MRG, a hierarchical heterogeneous graph-based medical report generation framework. In our framework, an anatomy detection module is adopted to extract fine-grained image patches of complete anatomical structures, replacing coarse fixed-size patches and enabling anatomy-aware visual tokenization. To mitigate false negatives in CLIP-style contrastive pretraining and enable precise visual-text alignment at the patch-sentence level, an anatomy-level contrastive sampling strategy is presented. More importantly, a new H^2-GAT is introduced, which explicitly models the structured reasoning pathway from anatomical evidence to disease inference through spatial, diagnostic, and co-occurrence relations, bridging anatomy-aware perception and clinically reliable report generation. Experiments on MIMIC-CXR and IU-Xray demonstrate state-of-the-art performance in both NLG and CE metrics. Future work will incorporate additional node types and relations for more comprehensive structured disease reasoning.

Limitations

This work has several limitations. First, H^2-MRG depends on an anatomy detection module to extract anatomy-aware visual tokens. Although our detector achieves strong localization performance, inaccurate or missing detections, especially for small anatomical structures, may still affect downstream disease reasoning and report generation. Second, our experiments focus on chest X-ray report generation, where anatomical regions and disease categories are defined using Chest ImaGenome and CheXpert. Extending the framework to other modalities or anatomical sites would require modality-specific anatomical definitions and anatomy-disease relations.

References

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *ECCV*, volume 13696 of *Lecture Notes in Computer Science*, pages 1–21.

Lifeng Chen, Tianqi You, Hao Liu, Zhimin Bao, Jile Jiao, Xiao Han, Zhicai Ou, Tao Sun, Xiaofeng Mou, Xiaojie Jin, and Yi Xu. 2026. ECHO: Efficient chest x-ray report generation with one-step block diffusion. *arXiv preprint arXiv:2604.09450*.

Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Guanbin Li, Cheng-Lin Liu, and Liang Lin. 2025. Cross-modal causal representation learning for radiology report generation. *IEEE Transactions on Image Processing*, 34:2970–2985.

Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In *ACL*, pages 9494–9509.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *ACL*, pages 5904–5914.

Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison O'Neil. 2023. Controllable chest X-ray report generation from longitudinal representations. In *Findings of EMNLP*, pages 4891–4904.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Yan Hu, Wenjie Li, and Jiang Liu. 2024. ICON: Improving inter-report consistency in radiology report generation via lesion-aware mixup augmentation. In *Findings of EMNLP*, pages 9043–9056.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Heng Li, Yan Hu, Wenjie Li, and Jiang Liu. 2025. RADAR: Enhancing radiology report generation with supplementary knowledge injection. In *ACL*, pages 26366–26381.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. 2023a. RECAP: Towards precise radiology report generation via dynamic disease progression reasoning. In *Findings of EMNLP*, pages 2134–2147.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023b. ORGAN: Observation-guided radiology report generation via tree reasoning. In *ACL*, pages 8108–8122.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. 2021. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *ICCV*, pages 3942–3951.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, volume 33, pages 590–597.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. PromptMRG: Diagnosis-driven prompts for medical report generation. In *AAAI*, volume 38, pages 2607–2615.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317.

Hanbin Ko and Chang-Min Park. 2025. Bringing CLIP to the clinic: Dynamic soft labels and negation-aware learning for medical analysis. In *CVPR*, pages 25897–25906.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023a. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *CVPR*, pages 3334–3343.

Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023b. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *ICCV*, pages 2863–2874.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024. Bootstrapping large language models for radiology report generation. In *AAAI*, volume 38, pages 18635–18643.

Haoxiang Liu, Sijun Bao, Shugeng Zhang, Jiancheng Wu, Chenhong Cao, and Wei Gong. 2026. Causal inference for reliable chest x-ray report generation. *Knowledge-Based Systems*, 337:115377.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. 2025. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *CVPR*, pages 10348–10359.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *NAACL*, pages 5288–5304.

Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2024. ChEX: Interactive localization and region description in chest x-rays. In *ECCV*, volume 15079 of *Lecture Notes in Computer Science*, pages 92–111.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. 2022. Joint learning of localized representations from medical images and reports. In *ECCV*, volume 13686 of *Lecture Notes in Computer Science*, pages 685–701.

Sang-Jun Park, Keun-Soo Heo, Dong-Hee Shin, Young-Han Son, Ji-Hye Oh, and Tae-Eui Kam. 2025. Dart: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In *CVPR*, pages 15580–15589.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, volume 28.

Mounir Salhi and Moulay A. Akhloufi. 2026. Recent progress in deep learning for chest x-ray report generation. *BioMedInformatics*, 6:3.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *CVPR*, pages 7433–7442.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2018. Graph attention networks. In *ICLR*.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022a. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *NeurIPS*, volume 35.

Jun Wang, Abhir Bhalerao, and Yulan He. 2022b. Cross-modal prototype driven network for radiology report generation. In *ECCV*, pages 563–579.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *CVPR*, pages 11558–11567.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022c. MedCLIP: Contrastive learning from unpaired medical images and text. In *EMNLP*, pages 3876–3887.

Joy T Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Alexander Paguio, Jasper Seth Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, and Mehdi Moradi. 2021. Chest ImaGenome dataset for clinical reasoning. In *NeurIPS Datasets Benchmarks*, volume 1.

Puzhen Wu, Hexin Dong, Yi Lin, Yihao Ding, and Yifan Peng. 2026. A disease-aware dual-stage framework for chest x-ray report generation. In *AAAI*, volume 40, pages 33953–33961.

Ting Xiao, Lei Shi, Yang Zhang, HaoFeng Yang, Zhe Wang, and Chenjia Bai. 2025. Online iterative self-alignment for radiology report generation. In *ACL*, pages 27799–27814.

Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80:102510.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. AlignTransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *MICCAI*, pages 72–82.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 2–25.

Yuanpin Zhou and Huogen Wang. 2024. Divide and conquer radiology report generation via observation level fine-grained pretraining and prompt tuning. In *EMNLP*, pages 7597–7610.

A Implementation Details

Our H^2-MRG follows a three-stage training pipeline.

Stage 1 (Anatomy Detection Training): The anatomy detector adopts Faster R-CNN with ResNet-50-FPN backbone and outputs 768-dimensional ROI features for 29 anatomical regions, with learned padding embeddings for missing regions. It is trained for 50 epochs using AdamW optimizer and ReduceLROnPlateau scheduler (learning rate 5×10^{-4}, batch size 128, weight decay 10^{-3}). The trained detector is frozen in the subsequent stages.

Stage 2 (Contrastive Pre-training): The ViT encoder is initialized from pretrained ViT Transformer weights (Dosovitskiy et al., 2021) and takes the detector outputs as inputs. A learnable positional embedding is assigned to each anatomical region and added to the corresponding input token. The encoder is pre-trained for 30 epochs using anatomy-level sampling and symmetric InfoNCE (Radford et al., 2021) (temperature $\tau = 0.07$, learning rate 10^{-4}, batch size 32).

Stage 3 (Supervised Fine-tuning): The report generator uses a GPT-2-based autoregressive Transformer decoder with cross-attention, initialized from pretrained GPT-2 weights. The cross-attention layers, H^2-GAT, and the disease classifier are newly initialized. Keeping the anatomy detector frozen, we jointly fine-tune the remaining pipeline, including the ViT encoder, cross-attention layers, H^2-GAT, disease classifier, and decoder, for 30 epochs (learning rate 5×10^{-5}, training/validation batch sizes 64/32, weight decay 0.001, gradient clipping norm 1.0, mixed precision FP16). The final objective combines cross-entropy loss $\mathcal{L}_{\text{CE}}$ for text generation and multi-label BCE loss $\mathcal{L}_{\text{BCE}}$ for disease classification with weight 0.1. Inference uses deterministic beam search with 3 beams. All experiments are conducted on a single NVIDIA A100 80GB GPU.

Anatomy-sentence and state extraction. Chest ImaGenome provides anatomy-sentence associations and normal/abnormal states for MIMIC-CXR through a rule-based NLP pipeline (Wu et al., 2021). Sentences are retained from the Findings and Impression sections, and anatomy mentions and attributes are linked within each sentence using named-entity recognition, syntactic parsing, and the dataset's radiologist-defined ontology. Affirmed and negated attribute relations are then grouped into the normal/abnormal states used by our sampling procedure. These automatically derived labels are used only for training. Errors in the extraction pipeline can introduce noise into anatomy-level contrastive pretraining, which we acknowledge in the Limitations section.

Reproducibility. We will release the implementation, preprocessing scripts, and graph-construction details needed to reproduce the training and evaluation pipeline from the public MIMIC-CXR and Chest ImaGenome releases.

Run-to-run variation. Across five independent runs on MIMIC-CXR, H^2-MRG obtains mean $\pm$ standard deviation scores of 0.436 ± 0.004, 0.293 ± 0.002, 0.210 ± 0.002, 0.160 ± 0.003, 0.186 ± 0.002, and 0.336 ± 0.004 for BLEU-1 through BLEU-4, METEOR, and ROUGE-L, respectively. The corresponding clinical-efficacy Precision, Recall, and F1 scores are 0.554 ± 0.004, 0.545 ± 0.004, and 0.547 ± 0.004.

Table 4: Comparison of graph operators on the same anatomy-disease graph on MIMIC-CXR.

Graph operator	**CE F1** $\uparrow$
GCN	0.513
Homogeneous GAT	0.528
H^2-GAT	**0.549**

B Alternative Graph Operators

To isolate the effect of relation-specific aggregation, we replace H^2-GAT with a homogeneous GAT or GCN while retaining the same anatomy and disease nodes, graph edges, input features, decoder, and training protocol. As shown in Table 4, H^2-GAT yields a clinical-efficacy F1 of 0.549, compared with 0.528 for the homogeneous GAT and 0.513 for the GCN. This controlled comparison indicates that typed relation-specific aggregation contributes beyond applying a generic graph operator to the same graph.

C Disease-Node Error Analysis

Disease nodes are latent representations rather than predicted labels directly injected into the decoder, and the decoder also retains a separate pathway to anatomy features. We nevertheless test whether disease-node errors are associated with unsupported generated entities using a reference-based hallucination proxy: the percentage of enti-

Table 5: Reference-based entity hallucination proxy on MIMIC-CXR.

Configuration	Rate ↓
ViT baseline	52.6%
w/o auxiliary classifier	45.7%
w/o $\mathcal{E}_{DD}$	44.5%
H^2-MRG	**42.3%**

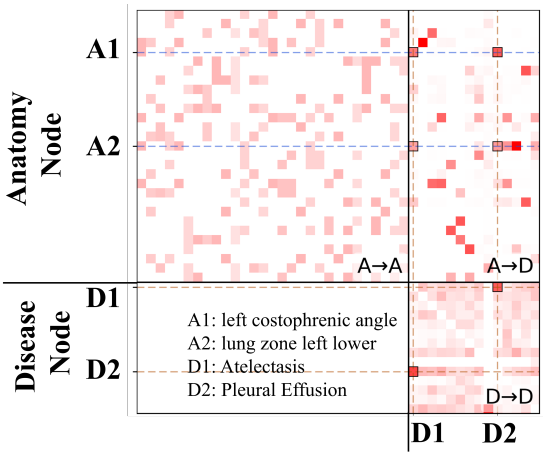


Figure 5: Node-level attention matrix in H^2-GAT for a case with comorbid Atelectasis and Pleural Effusion. A1 and A2 present two anatomy nodes, while D1 and D2 are two disease nodes.

ties extracted from a generated report that have no matched entity with a compatible clinical status in the reference report. Lower values are better.

On the subset for which the auxiliary classifier makes at least one error, the proxy rate is 51.8% for H^2-MRG and 57.4% for the ViT baseline. These results provide no evidence that the proposed disease representations increase this reference-based proxy, including on classifier-error cases. Because a reference report may itself omit valid observations, this proxy should not be interpreted as a clinical adjudication of hallucination.

D H^2-GAT Attention Visualization

We visualize the node-level attention matrix of H^2-GAT for a case with comorbid Atelectasis and Pleural Effusion in Fig. 5. The high responses on disease-associated anatomy and disease-disease interaction regions show that disease predictions are grounded in anatomically relevant evidence and refined by co-occurrence priors.

E Sensitivity to Anatomy Detection

We assess the effect of anatomy detection accuracy by perturbing ground-truth bounding boxes

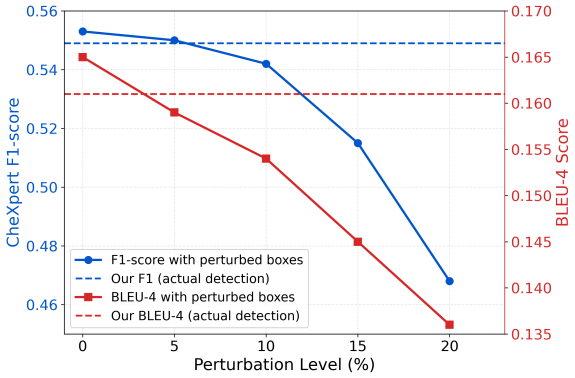


Figure 6: CheXpert F1 (left, blue) and BLEU-4 (right, red) under different detection perturbation levels. Dashed lines indicate our trained detector's performance.

at 5%, 10%, 15%, and 20% levels (random center shifts and scale jittering). As shown in Fig. 6, performance degrades smoothly: at 5% perturbation, F1 remains 0.550 (vs. 0.553 with GT boxes) and BLEU-4 remains 0.159 (vs. 0.165). Degradation becomes more apparent at 20% (F1 0.468, BLEU-4 0.136). With the trained detector, our system reaches F1 0.549 and BLEU-4 0.161, close to the GT-box upper bound, confirming practical robustness.

F Contrastive Learning Analysis

To validate the anatomy-level sampling strategy, we compare our approach against image-report-level CLIP (IR-CLIP) (Radford et al., 2021) and patch-sentence-level CLIP (PS-CLIP) within the same framework. IR-CLIP aligns entire images with reports, while PS-CLIP aligns patches with sentences but treats all unpaired samples as negatives. As shown in Table 7, IR-CLIP achieves the lowest performance due to the lack of fine-grained alignment, while PS-CLIP improves but still suffers from false negatives. Our anatomy-level sampling strategy achieves the highest performance, confirming that cross-patient anatomical grouping improves visual-text alignment.

In addition, we analyze the scale of contrastive pairs (Table 8). Our anatomy-level sampling enriches positives to 4,018 (vs. 32 for IR-CLIP and 301 for PS-CLIP) through cross-sample anatomical grouping while maintaining a large negative pool of 86,513, providing dense and semantically clean supervision.

Anatomical Region	Mean IoU	Precision@0.5	Recall@0.5	F1@0.5	AP@0.5	Missing Rate
Left Upper Lung Zone	0.9030	0.9705	0.9746	0.9725	0.9065	0.0036
Abdomen	0.9012	0.9801	0.9927	0.9864	0.9090	0.0007
Left Lung	0.8995	0.9830	0.9911	0.9870	0.9086	0.0007
Right Lung	0.8971	0.9842	0.9888	0.9865	0.9073	0.0000
Right Upper Lung Zone	0.8922	0.9678	0.9739	0.9709	0.9066	0.0010
Right Lower Lung Zone	0.8813	0.8634	0.9116	0.8868	0.8818	0.0147
Spine	0.8736	0.9888	0.9918	0.9903	0.9091	0.0023
Left Lower Lung Zone	0.8725	0.8744	0.9196	0.8964	0.8790	0.0185
Mediastinum	0.8620	0.9404	0.9825	0.9610	0.9047	0.0000
Left Mid Lung Zone	0.8575	0.9042	0.9455	0.9244	0.8857	0.0055
Right Mid Lung Zone	0.8571	0.8824	0.9346	0.9078	0.8874	0.0068
Left Apical Zone	0.8570	0.9240	0.9495	0.9366	0.9011	0.0072
Right Hilar Structures	0.8567	0.8910	0.9347	0.9123	0.8892	0.0048
Right Clavicle	0.8518	0.8868	0.8809	0.8838	0.8113	0.0445
Upper Mediastinum	0.8510	0.9234	0.9505	0.9368	0.9018	0.0056
Left Hilar Structures	0.8497	0.9008	0.9400	0.9200	0.8818	0.0065
Right Apical Zone	0.8470	0.9196	0.9433	0.9313	0.8992	0.0058
Right Costophrenic Angle	0.8446	0.8076	0.7593	0.7827	0.6986	0.0766
Aortic Arch	0.8434	0.9228	0.8984	0.9104	0.8096	0.0355
Cardiac Silhouette	0.8428	0.8985	0.9542	0.9255	0.8968	0.0043
Right Hemidiaphragm	0.8359	0.8202	0.8524	0.8360	0.7937	0.0179
Left Clavicle	0.8331	0.8686	0.8749	0.8717	0.8084	0.0372
Left Costophrenic Angle	0.8218	0.7841	0.7612	0.7725	0.6976	0.0693
SVC	0.8168	0.8845	0.8737	0.8791	0.7955	0.0202
Left Hemidiaphragm	0.8108	0.8542	0.8858	0.8697	0.7970	0.0123
Trachea	0.7919	0.9318	0.9330	0.9324	0.8985	0.0086
Right Atrium	0.7756	0.8862	0.8713	0.8787	0.8045	0.0301
Cavoatrial Junction	0.7341	0.7469	0.7249	0.7357	0.6585	0.0461
Carina	0.7043	0.6951	0.6027	0.6456	0.5527	0.1444
Average	**0.8436**	**0.8940**	**0.9034**	**0.8976**	**0.8407**	**0.0218**

Table 6: Detailed anatomy detection results for all 29 anatomical regions ranked by mean IoU. Precision, Recall, F1, and AP are computed at IoU threshold 0.5. Missing Rate indicates the proportion of images where the anatomical region was not detected.

Table 7: Contrastive pretraining results on MIMIC-CXR.

Metric	IR-CLIP	PS-CLIP	Ours
BLEU-1 ↑	0.407	0.429	**0.434**
BLEU-2 ↑	0.269	0.290	**0.295**
BLEU-3 ↑	0.184	0.202	**0.208**
BLEU-4 ↑	0.134	0.148	**0.161**
METEOR ↑	0.159	0.181	**0.185**
ROUGE-L ↑	0.317	0.334	**0.338**
CE Precision ↑	0.506	0.536	**0.556**
CE Recall ↑	0.494	0.524	**0.543**
CE F1 ↑	0.500	0.530	**0.549**

Strategy	Positives	Negatives
IR-CLIP	32	992
PS-CLIP	301	84,662
Ours (Anatomy-level)	4,018	86,513

Table 8: Contrastive-pair counts with batch size 32.

G Detailed Anatomy Detection Results

Table 6 reports detection results for the 29 Chest ImaGenome regions. The detector reaches a mean IoU of 0.8436, an F1-score of 0.8976, and an average missing rate of 2.18%. Lungs and abdomen exceed 0.87 IoU and 0.98 F1, whereas small structures such as the carina (IoU=0.7043) and cavoatrial junction (IoU=0.7341) remain more challenging because of their size and boundary complexity.