FAMDR: Feature-Aligned Multimodal Denoising for Reliable Diagnostic Reconciliation in Medical Imaging

Xun Liang*, Zhiying Li*, Hongxun Jiang†

School of Information, Renmin University of China
No. 59 Zhongguancun Street, Beijing, 100872, China
{xliang, zhiyingli, jianghx}@ruc.edu.cn

Abstract

This paper presents FAMDR, a Feature-Aligned Multimodal Denoising framework for Reliable Diagnostic Reconciliation. Existing approaches suffer from two major limitations: (1) an overemphasis on simplifying observational descriptions and (2) a failure to denoise the misleading content in radiological findings against clinical histories. Current methods often dismiss such cross-modal inconsistencies as noise rather than clinically significant signals. To bridge this gap, the framework integrates four synergistic components: (1) noise-aware multimodal alignment that preserves discriminative discrepancy features while ensuring semantic coherence, (2) cross-modal retrieval augmentation leveraging external medical knowledge to resolve ambiguous cases, (3) granular localization of noises at pixel and phrase levels using adaptive thresholding, and (4) medical noise uncertainty quantification to provide reliable confidence estimates. Evaluated on an extended MIMIC-CXR dataset enriched with expert-annotated noise and longitudinal records, FAMDR achieves superior accuracy in semantic denoising and inconsistency localization while preserving clinical interpretability. Its capability to generate actionable, uncertainty-aware reports advances safer and more reliable integration into diagnostic workflows.

Introduction

Automated diagnostic report generation has revolutionized clinical workflows—dramatically reducing radiologist burden and enhancing diagnostic efficiency—yet it remains prone to introducing subtle noise and misleading content that can compromise clinical precision. As shown in Figure 1, a 50-year-old male with fever, night sweats, and cough was initially diagnosed via chest X-ray with left pleural effusion and lower lung opacity, suggestive of community-acquired pleuropneumonia. However, antibiotics failed, and a deeper review revealed a splenectomy 34 years earlier following traumatic hemothorax and diaphragmatic rupture. This case underscores the challenge of reconciling imaging findings with clinical history, especially when textual reports misalign with radiological observations (Tulinský et al. 2016). Manually cross-referencing extensive patient histories remains impractical for radiologists. Automated real-time semantic denoising enhances the accuracy, consistency, and trustworthiness of reports by detecting clinical-logical discrepancies between medical imaging findings and electronic health records (EHRs).

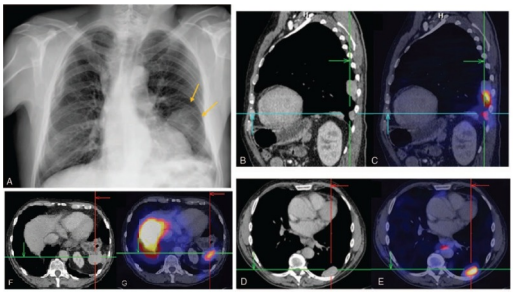

Figure 1: Illustration of multimodal denoising between imaging and pathology in thoracic splenosis mimicking pleuropneumonia. Chest X-ray (A) showed left-sided opacity. Fused CT (B, D, F) and Tc-99m colloid SPECT (C, E, G) revealed hypervascular thoracic and abdominal masses with intense uptake, confirming ectopic splenic tissue. Benign postoperative finding, no treatment needed.

Recent advances in AI for radiology and EHRs have enabled vision-text alignment (Dawidowicz, Hirsch, and Tal 2023; Wang et al. 2022), report generation (Xu et al. 2023; Tanida et al. 2023), and joint representation learning (Chen et al. 2024). Yet, these methods lack explicit mechanisms for detecting factual contradictions between imaging and clinical narratives, often dismissing cross-modal inconsistencies as noise. This undermines their reliability in real-world clinical settings, where interpretive noise in medical imaging may indicate potential diagnostic risks.

We propose FAMDR to deliver interpretations that are both robust and clinically coherent. FAMDR models and resolves medical inconsistencies between radiology images and EHRs via: (a) dynamic discrepancy modeling with learnable discrepancy vectors, (b) evidence-aware reasoning through real-time knowledge retrieval and adaptive gating, (c) threshold-free, anatomy-agnostic inconsistency localization, and (d) composite uncertainty calibration spanning data, knowledge, and model dimensions. Unlike binary classifiers, FAMDR emulates clinical reasoning to sur-

*These authors contributed equally.
†Corresponding author.

face contradictions while preserving report completeness, advancing precision denoising aligned with medical science and clinical standards.

Related Works

Research in medical report generation has progressed from early single-modality encoder-decoder models to more sophisticated multimodal fusion systems. Initial approaches like R2Gen (Chen et al. 2020) and CMN (Chen et al. 2021) adapted image captioning techniques but struggled with generating detailed, long-form clinical text. To improve semantic representation, models such as KiUT (Huang, Zhang, and Zhang 2023) and DCL (Li et al. 2023) incorporated domain knowledge via static semantic graphs, which limited adaptability to new information. CoE-DG (Sun et al. 2024) introduced a co-evolutionary design for abnormality detection and report generation using bidirectional interactions (GIP/DIP), yet failed to address multimodal contradictions. PromptMRG (Jin et al. 2024) leveraged diagnosis-oriented prompts and CLIP-based case retrieval to enhance clinical accuracy but assumed complete diagnostic labels and overlooked discrepancies between medical history and imaging reports.

In cross-modal alignment, BioMedCLIP (Zhang et al. 2023) and XrayGPT (Thawkar et al. 2023) explored joint vision-language representations, but their general-purpose training hindered performance on clinical contradiction detection. CXRmate (Nicolson et al. 2024) and Recap (Hou et al. 2023) incorporated longitudinal imaging data to model disease progression, but did not explicitly reconcile inconsistencies across textual modalities, risking historical noise. RadFusion (Mei et al. 2024) introduced spatio-temporal fusion with feature normalization to handle imbalanced data but lacked fine-grained contradiction modeling in text.

Trust and uncertainty remain underexplored. M2factENTNL (Miura et al. 2021) optimized factual accuracy via reinforcement learning but showed inconsistent clinical performance. PromptMRG's (Jin et al. 2024) adaptive logit adjustment improved class balance but ignored uncertainty quantification. CoE-DG's (Sun et al. 2024) pseudo-label refinement filtered noise using static thresholds, limiting confidence calibration under complex and inconsistent scenarios.

Methodology

Figure 2 presents the proposed FAMDR framework, which delivers clinically reliable decision support through four synergistic components. The core task is defined as follows: given a medical image $I \in \mathbb{R}^{H\times W\times C}$ and its associated EHRs $T = \{t_1, t_2, ..., t_n\}$, the goal is to learn a mapping $F : (I, T) \rightarrow (R, K, D, c)$, where $R \subseteq I$ and $K \subseteq T$ identify contradictory regions in the images and texts, D denotes diagnostic identifications, and $c \in [0, 1]$ is a calibrated confidence score. An inconsistency is flagged when $\text{Sim}(f_I(R), f_T(K)) < \delta$ with f_I and f_T presenting modality-specific encoders and δ is a dynamic threshold learned during training, initialized at 0.4 and optimized via CAMAL's discrepancy regularization loss.

we utilize an image encoder, ViT (Dosovitskiy et al. 2020), and a text encoder, ClinicalBERT (Huang, Altosaar, and Ranganath 2019), to map I and T into the raw visual and textual feature spaces. In detail, for an image-record pair (I, T), the image encoder generates a sequence of encoded visual tokens $V = \{v_i\}_{i=1}^{N}$ ($v_i \in \mathbb{R}^{d_v}$) and a global image representation v. Similarly, the text encoder generates a sequence of encoded text tokens $T = \{t_j\}_{j=1}^{M}$ ($t_j \in \mathbb{R}^{d_t}$) and a global record representation t. Here, N and M denote the total number of visual and textual tokens, respectively.

CAMAL: Discrepancy-Aware Multimodal Alignment

Conventional multimodal alignment overlooks discrepancy modeling, limiting effectiveness in medical contradiction detection. We present a dual-stream framework that jointly learns modality-invariant features and quantifies discrepancies through learnable discrepancies, preserving critical signals while ensuring semantic coherence.

We first project the raw visual and textual tokens into a shared d-dimensional space, $v'_i = W_v v_i + b_v$ and $t'_j = W_t t_j + b_t$, where $W_v \in \mathbb{R}^{d\times d_v}$, $W_t \in \mathbb{R}^{d\times d_t}$ are learnable projection matrices. Discrepant feature vectors (e.g., a lung opacity in imaging vs. 'no consolidation' in text) must be aligned in a shared space to quantify their directional discrepancy while preserving semantic coherence. This enables distinguishing true contradictions from benign variations. For each token pair (v'_i, t'_j), we compute the discrepancy feature vector through normalized discrepancy: $c_{ij} = \phi\left(\frac{v'_i - t'_j}{\|v'_i - t'_j\|_2}\right) \in \mathbb{R}^d$, where $\phi(\cdot)$ denotes a GeLU-activated MLP. The cross-attention mechanism based discrepancy-aware alignment (Lai et al. 2024; Liu et al. 2024) then operates as $\hat{v}_i = \text{softmax}\left(\frac{c_{ij}(W_q^v[v'_1,...,v'_N])^\top}{\sqrt{d}}\right) V'$ for visual alignment and similarly for textual alignment, where W_q^v, W_q^t are query transformation matrices. $v^i \in \mathbb{R}^d$ represents a patch embedding from ViT, corresponding to a region of the input image (e.g., 16×16 pixels). The alignment loss combines contrastive learning with discrepancy regularization:

$$\mathcal{L}_{\text{align}} = \sum_{i,j} -\log \frac{e^{\text{sim}(\hat{v}_i,\hat{t}_j)/\tau}}{\sum_k e^{\text{sim}(\hat{v}_i,\hat{t}_k)/\tau}} + \lambda \sum_{i,j} \|c_{ij} - (\hat{v}_i \odot \hat{t}_j)\|_2^2 \tag{1}$$

Here, $\text{sim}(\hat{v}_i, \hat{t}_j)$ computes token-wise cosine similarity, τ is a temperature scalar, and λ balances two objectives: (i) Cross-Modal Contrast pulls matching image-text pairs closer while pushing mismatched pairs apart; (ii) Discrepancy Consistency regularizes the discrepancy vector c_{ij} to align with the residual of aligned embeddings $\hat{v}_i \odot \hat{t}_j$, ensuring discrepancies reflect genuine contradictions rather than noise. Positive pairs $(\hat{v}_i, \hat{t}_j)$ are aligned image-text tokens; negatives are in-batch non-matching pairs $(\hat{v}_i, \hat{t}_{k\neq j})$.

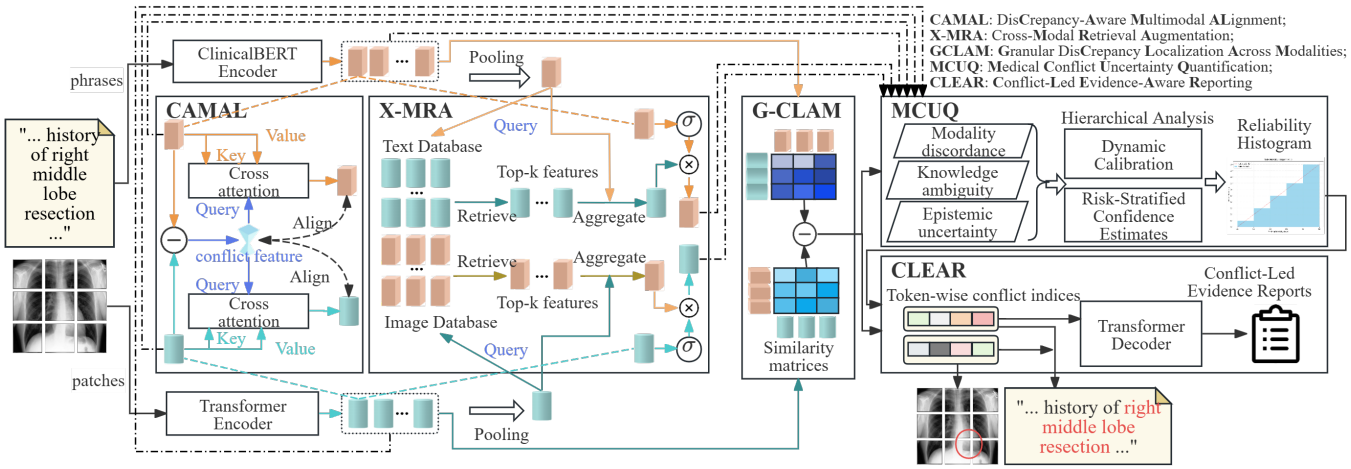

Figure 2: The FAMDR framework comprises five modules—CAMAL, X-MRA, G-CLAM, MCUQ, and CLEAR—designed for clinically reliable medical inconsistency identification. CAMAL learns modality-invariant features and models discrepancies via discrepancy-aware cross-attention. X-MRA integrates external medical knowledge through gated fusion based on the likelihood of inconsistency. G-CLAM localizes contradictions at pixel and phrase levels using adaptive thresholds without relying on predefined lexicons. MCUQ provides calibrated uncertainty estimates by decomposing data, knowledge, and model uncertainty into risk-stratified confidence scores. CLEAR explicitly highlights multimodal noises when generating radiological imaging reporting.

X-MRA: cross-Modal Retrieval Augmentation

Existing medical vision-language models often lack clinical context due to limited data. We propose a retrieval-enhanced alignment paradigm that enriches text with radiology lexicons and grounds visuals in imaging biomarkers. A gated fusion mechanism adaptively balances inputs with retrieved evidence based on discrepancy likelihood.

Let $V^{\text{orig}} = \{v_i\}_{i=1}^{N} \in \mathbb{R}^{N\times d}$ and $T^{\text{orig}} = \{t_j\}_{j=1}^{M} \in \mathbb{R}^{M\times d}$ denote the projected visual and textual token embeddings. We first derive global representations through masked average pooling, $v_g = \frac{1}{\sum_{i=1}^{N} m_i^v}\sum_{i=1}^{N} m_i^v v_i$ and t_g. Discrepancy masks $m_i^v \in \{0,1\}$ and $m_j^t \in \{0,1\}$ are derived by thresholding the discrepancy vector magnitude $\|c_{ij}\|_2$. Tokens with $\|c_{ij}\|_2 > \eta$ (where η is the 75%-ile of magnitudes) are masked ($m = 0$) to suppress high-discrepancy features during global pooling. The retrieval process operates bidirectionally: $\mathcal{R}_t = \text{TopK}\,(\text{sim}(t_g, \mathcal{D}_t))$, $\hat{t}_j = t_j + \gamma_t \cdot \text{Attn}(t_j, \mathcal{R}_t)$ and similarly for $\mathcal{R}_v$ and $\hat{v}_i$. Here, $\mathcal{D}_t$ and $\mathcal{D}_v$ are pretrained medical text/image databases, $\text{Attn}(q, \mathcal{R}) = \sum_{k=1}^{K} \text{softmax}(q^\top r_k/\tau) r_k$ performs attention-based aggregation, and $\gamma_{t/v} \in [0,1]$ are trainable gating parameters computed as $\gamma_t = \sigma(W_t[t_j; \text{mean}(\mathcal{R}_t)])$ and $\gamma_v = \sigma(W_v[v_i; \text{mean}(\mathcal{R}_v)])$. The final loss incorporates retrieval consistency:

$$\mathcal{L}_{\text{ret}} = \|\hat{V} - V\|_{\text{FRO}} + \|\hat{T} - T\|_{\text{FRO}} + \beta \sum_{k=1}^{K} \text{sim}(v_g, r_k^v) \cdot \text{sim}(t_g, r_k^t) \tag{2}$$

$\mathcal{L}_{ret}$ ensures: (i) Visual/Textual Fidelity ($\|\hat{V} - V\|_{\text{FRO}}$, $\|\hat{T} - T\|_{\text{FRO}}$) retains original semantics after retrieval; (ii) Cross-Modal Consistency ($\sum \text{sim}(v_g, r_v^k) \cdot \text{sim}(t_g, r_t^k)$) aligns retrieved image/text evidence to resolve ambiguities (e.g., differentiating tumor vs. scarring). Unlike standard retrieval, X-MRA uses discrepancy likelihood $\gamma_{t/v}$ to gate irrelevant knowledge, preventing noise amplification.

G-CLAM: Granular disCrepancy Localization Across Modalities

Accurate medical semantic denoising requires multi-evidence fusion and discrepancy-aware reasoning across hierarchical features. Our framework introduces: (a) tri-source feature integration preserving original inputs while leveraging aligned and retrieved knowledge, (b) bidirectional similarity analysis with directional discrepancy quantification, and (c) threshold-free discrepancy indexing via adaptive residual learning. This enables pixel/phrase-level localization without predefined anatomical lexicons.

Let V^{orig}, V^{align}, V^{ret} denote the original, aligned, and retrieval-augmented visual embeddings ($\in \mathbb{R}^{N\times d}$), with corresponding textual embeddings T^{orig}, T^{align}, T^{ret} ($\in \mathbb{R}^{M\times d}$). We first fuse multi-source features through residual-aware summation, $\hat{v}_i = v_i^{\text{orig}} + v_i^{\text{align}} + v_i^{\text{ret}}$ and $\hat{t}_j = t_j^{\text{orig}} + t_j^{\text{align}} + t_j^{\text{ret}}$. Bidirectional similarity matrices are computed as: $S_{v2t}[i,j] = \frac{\hat{v}_i^\top \hat{t}_j}{\|\hat{v}_i\|\|\hat{t}_j\|}$, $S_{t2v}[j,i] = \frac{\hat{t}_j^\top \hat{v}_i}{\|\hat{t}_j\|\|\hat{v}_i\|}$, $\Delta S = |S_{v2t} - S_{t2v}^\top| \odot M_{\text{discrepancy}}$. Here, $M_{\text{discrepancy}} \in \{0,1\}^{N\times M}$ masks non-conflicting pairs using CAMAL's outputs. The discrepancy indices are determined by $\mathcal{C}_{\text{v}} = \{i | \sum_{j=1}^{M} \Delta S[i,j] > \tau_v\}$ and $\mathcal{C}_{\text{t}} = \{j | \sum_{i=1}^{N} \Delta S[i,j] > \tau_t\}$ with adaptive thresholds learned via $\tau_v = \sigma\left(W_{\tau v}[\text{mean}(\Delta S); \text{std}(\Delta S)]\right)$ and $\tau_t = \sigma\left(W_{\tau t}[\text{mean}(\Delta S^\top); \text{std}(\Delta S^\top)]\right)$. G-CLAM's adaptive threshold τ_v, τ_t uses sigmoid activations to dynamically separate inconsistent tokens based on localized discrepancy

density, avoiding fixed anatomical lexicons. The localization loss combines ranking optimization with spatial consistency:

$$\mathcal{L}_{\text{loc}} = \sum_{i \in \mathcal{C}_{\text{v}}} \sum_{j \in \mathcal{C}_{\text{t}}} \max(0, \alpha - (\Delta S[i,j] - \mathbb{E}[\Delta S])) + \beta \|\nabla_{\text{s}}(\Delta S)\|_1 \tag{3}$$

where α controls margin width and ∇_{s} enforces smoothness in anatomically adjacent regions.

MCUQ: Medical Conflict Uncertainty Quantification

Medical conflict uncertainty arises from three sources: (a) modality discordance between images and text, (b) ambiguity in retrieved clinical knowledge, and (c) epistemic uncertainty due to limited data. FAMDR addresses these via hierarchical uncertainty decomposition and dynamic calibration to produce calibrated confidence estimates. This design aligns with Bayesian perspectives, mapping data noise, retrieval ambiguity, and model variability to distinct uncertainty types. Our key innovation disentangles epistemic and aleatoric uncertainty while maintaining clinical interpretability.

Let $\mathcal{F} = \{V^{\text{align}}, T^{\text{align}}, \mathcal{C}_v, \mathcal{C}_t, \mathcal{R}_v, \mathcal{R}_t\}$ denote inputs from preceding modules. We define the uncertainty score $U \in [0,1]$ through multi-component aggregation: $U = \omega_1 \cdot \mathcal{U}_{\text{mod}} + \omega_2 \cdot \mathcal{U}_{\text{know}} + \omega_3 \cdot \mathcal{U}_{\text{model}}$. Here, **Modality Discordance** quantifies feature-space discrepancies: $\mathcal{U}_{\text{mod}} = 1 - \frac{1}{|\mathcal{C}_v||\mathcal{C}_t|} \sum_{i \in \mathcal{C}_v} \sum_{j \in \mathcal{C}_t} \frac{v_i^{\text{align}\top} t_j^{\text{align}}}{\|v_i^{\text{align}}\| \|t_j^{\text{align}}\|}$, **Knowledge Ambiguity** evaluates retrieval consistency: $\mathcal{U}_{\text{know}} = \frac{1}{K} \sum_{k=1}^{K} \left(1 - \frac{\text{sim}(\mathcal{R}_v^k, \mathcal{R}_t^k)}{\max(\text{sim}(\mathcal{R}_v^k, T^{\text{align}}), \text{sim}(\mathcal{R}_t^k, V^{\text{align}}))}\right)$, and **Epistemic Uncertainty** measures model confidence via Monte Carlo dropout (Gal and Ghahramani 2016; Lemay et al. 2022): $\mathcal{U}_{\text{model}} = \frac{1}{S} \sum_{s=1}^{S} \|f_{\theta_s}(I,T) - \mu_{\text{pred}}\|_2$, $\mu_{\text{pred}} = \frac{1}{S} \sum_{s=1}^{S} f_{\theta_s}(I,T)$, where θ_s denotes parameters with dropout masks applied during $S = 10$ stochastic forward passes. The adaptive weights of three uncertainties ω_i are dynamically calibrated through $\omega_i = \exp(W_i[\mathcal{U}_{\text{mod}}; \mathcal{U}_{\text{know}}; \mathcal{U}_{\text{model}}]) / \sum_{j=1}^{3} \exp(W_j[\mathcal{U}_{\text{mod}}; \mathcal{U}_{\text{know}}; \mathcal{U}_{\text{model}}])$.

Calibration Loss (Krishnan and Tickoo 2020) ensures uncertainty estimates align with empirical error rates $\mathcal{L}_{\text{cal}} = \text{MMD}(p(U|\text{conflict}), p(U|\text{non-conflict})) + \text{ECE}(U, \mathbb{I}_{\text{error}})$, where MMD quantifies distribution divergence and ECE measures calibration error (Gruber and Buettner 2022).

CLEAR: Conflict-Led Evidence-Aware Reporting

Conventional medical report generation lacks explicit denoising and clinical coherence. CLEAR addresses this with: 1) discrepant-attentive decoding to highlight contradictions, 2) retrieval-augmented identifications grounded in guidelines, and 3) structured templates aligned with standards (e.g., BI-RADS). This ensures precise, interpretable documentation of contradictions for clinical decision-making.

Let $\mathcal{C}_v = \{i_1, ..., i_K\}$ and $\mathcal{C}_t = \{j_1, ..., j_L\}$ denote the visual/textual discrepancy indices from G-CLAM. The decoder first constructs discrepancy-aware attention masks $M \in \{0, -\infty\}^{N \times M}$: $M[i,j] = \begin{cases} 0 & \text{if } i \in \mathcal{C}_v \text{ and } j \in \mathcal{C}_t \\ -\infty & \text{otherwise} \end{cases}$. The transformer decoder generates tokens y_t conditioned on: $p(y_t|y_{<t}, I, T) = \text{softmax}\left(E^\top h_t^{\text{dec}}\right)$, where h_t^{dec} is computed through discrepancy-guided cross-attention: $h_t^{\text{dec}} = \text{TransformerBlock}(Q = h_{t-1}^{\text{dec}}, K = V^{\text{align}}, V = V^{\text{align}} + \text{Attn}(M \odot (V^{\text{align}} T^{\text{align}\top})))$. Clinical recommendations are retrieved from guidelines $\mathcal{G}$ using the current context $c_t = \text{mean}(h_{1:t}^{\text{dec}})$: $\mathcal{R}_t = \text{TopK}(\text{sim}(c_t, \mathcal{G}))$. With the training objective fusing language modeling with retrieval alignment $\mathcal{L}_{\text{gen}} = \text{CE-Loss}(y_{1:T}, y^{\text{GT}}) + \lambda \sum_{k=1}^{K} \text{sim}(c_{t_k}, \mathcal{R}_t[k])$. The final output combines generated text with retrieved evidence: $\text{Report} = \{y_1, ..., y_T\} \oplus \bigcup_{k=1}^{K} \mathcal{R}_t[k]$.

Experiments

To rigorously assess the effectiveness of the FAMDR framework, we conduct comprehensive experiments guided by four research questions (RQs): **RQ1:** Does FAMDR outperform existing models in detecting and localizing multimodal medical inconsistencies? **RQ2:** How do FAMDR's core innovations individually contribute to its overall performance? **RQ3:** Can FAMDR generate clinically interpretable inconsistency traces that align with physician workflows while resolving real-world diagnostic ambiguities? **RQ4:** How does FAMDR dynamically prioritize multimodal evidence to emulate clinical reasoning under uncertainty? **RQ5:** How do key hyperparameters affect FAMDR's performance? Experimental settings can be found in Appendix.

RQ1: Main Comparison Results

Technical Innovation and Cross-Modal Alignment. Table 1 highlights FAMDR's strong performance over both general-purpose and medical models, particularly in semantic denoising (F1: 0.83 vs. 0.76 for MedKLIP) and localization (IoU: 0.59 vs. 0.45). While models like GLoRIA and MGCA rely on contrastive learning for hierarchical alignment, they lack explicit inconsistency modeling, limiting their ability to detect subtle contradictions (e.g., between imaging biomarkers and negations). MedKLIP and KiUT incorporate clinical knowledge but are restricted by static knowledge graphs. In contrast, FAMDR's dual-stream alignment and retrieval-augmented grounding enable adaptive, uncertainty-aware reasoning, delivering a 9% F1 gain over the best baseline.

Clinical Reliability and Granularity. Table 1 highlights FAMDR's superior discrepancy localization at the pixel/phrase level, with an IoU of 0.59 vs. 0.43 for MGCA. Models leveraging historical data (e.g., Recap, CXRmate) show limited spatial precision due to their focus on temporal patterns, while large vision-language models like XrayGPT and LLaVA-Med prioritize fluency over contradiction detection. In contrast, FAMDR's multi-source feature fusion and adaptive thresholding preserve critical inconsistencies, achieving clinician-level localization accuracy.

Uncertainty Quantification and Trustworthiness. FAMDR achieves superior uncertainty calibration (ECE: 0.08 vs. 0.10–0.18 for baselines), reinforcing its clinical

Model	F1 ↑	IoU ↑	ECE ↓
CLIP	0.62±0.03	0.28±0.02	0.19±0.02
GPV	0.65±0.02	0.31±0.03	0.17±0.01
GLIPv2	0.67±0.03	0.33±0.02	0.15±0.02
CMN	0.71±0.02	0.38±0.03	0.14±0.01
GLoRIA	0.73±0.03	0.41±0.02	0.13±0.02
MGCA	0.74±0.02	0.43±0.03	0.12±0.01
Recap	0.68±0.03	0.35±0.02	0.16±0.02
ORGan	0.72±0.02	0.39±0.03	0.14±0.01
KiUT	0.75±0.03	0.44±0.02	0.11±0.02
DCL	0.70±0.02	0.37±0.03	0.15±0.01
XrayGPT	0.69±0.03	0.34±0.02	0.18±0.02
LLaVA-Med	0.73±0.02	0.40±0.03	0.13±0.01
BioMedCLIP	0.70±0.03	0.36±0.02	0.16±0.02
MedKLIP	0.76±0.02	0.45±0.03	0.10±0.01
RadFusion	0.67±0.03	0.32±0.02	0.17±0.02
PromptMRG	0.74±0.02	0.42±0.03	0.12±0.01
CheXagent	0.71±0.03	0.38±0.02	0.14±0.02
CXRmate	0.66±0.02	0.30±0.03	0.18±0.01
FAMDR (**Ours**)	**0.83******±0.02**	**0.59******±0.03**	**0.08******±0.01**

Table 1: Performance comparison on denoising (F1), localization (IoU), and uncertainty calibration (ECE). Higher F1/IoU and lower ECE indicate better performance. $^{**}p < 0.01$ vs. best baseline (paired t-test).

reliability. General-purpose models like CLIP and GPV show higher ECE due to overconfidence in ambiguous cases, while knowledge-enhanced models (e.g., MedKLIP) struggle to separate modality discordance from epistemic uncertainty. In contrast, FAMDR's hierarchical uncertainty decomposition dynamically weights discordance, knowledge ambiguity, and model confidence, aligning risk estimates with empirical error rates.

RQ2: Ablation Study

The ablation study isolates the contribution of each core FAMDR module while preserving task fidelity. G-CLAM, responsible for final denoising and localization, cannot be removed and is instead modified by replacing adaptive residual learning with static cosine similarity to assess threshold-free adaptation. CLEAR is excluded, as it functions solely as a standardized post-processing module; ablating it would conflate generation with inconsistency detection. While CAMAL, X-MRA, and MCUQ could be replaced with alternatives (e.g., adversarial modeling, memory networks, Bayesian ensembles), we opt for full removal to evaluate core necessity over incremental variants. Due to the framework's modular design and clearly separated interfaces, inter-module dependencies do not confound results—each ablation reflects direct performance impact of the target module.

Explicit Discrepancy Modeling vs. Implicit Alignment. The sharp performance drop without CAMAL (-12% F1, -24% IoU) highlights the importance of explicit discrepancy pattern modeling, as shown in Table 2. Standard alignment approaches prioritize semantic coherence but miss directional discrepancies across modalities, such as negation

Model	F1 ↑	IoU ↑	ECE ↓
FAMDR	**0.83**	**0.59**	**0.08**
w/o CAMAL	0.73 (-12%)	0.45 (-24%)	0.11 (+38%)
w/o X-MRA	0.78 (-6%)	0.51 (-14%)	0.10 (+25%)
w/o MCUQ	0.80 (-4%)	0.56 (-5%)	0.12 (+50%)
G-CLAM → Cosine	0.75 (-10%)	0.49 (-17%)	0.10 (+25%)

Table 2: Ablation analysis of FAMDR components.

phrases contradicting imaging findings. CAMAL's dual-stream architecture quantifies such discrepancies using normalized discrepancy vectors, enabling clear separation between inconsistent and aligned pairs. This reduces over-smoothing in cross-attention, as shown by the 24% IoU gap between FAMDR and the w/o CAMAL.

Retrieval-Augmented Grounding and Uncertainty Synergy. Table 2 highlights the complementary functions of X-MRA and MCUQ. Removing X-MRA reduces detection performance (F1: –6%) and significantly worsens calibration (ECE: +25%), as the model loses access to external knowledge needed to resolve ambiguous cases (e.g., rare diseases). In contrast, disabling MCUQ mainly impacts calibration (ECE: +50%) with minimal F1 decline (–4%), showing that uncertainty modeling enhances trustworthiness without affecting base accuracy. Their synergy is most evident in complex scenarios—retrieval provides context, while uncertainty weighting dynamically modulates evidence credibility.

Adaptive Localization vs. Static Thresholds. Replacing G-CLAM's adaptive thresholds with cosine similarity leads to a larger drop in localization performance (–17% IoU vs. –14% F1), underscoring the difficulty of anatomically-agnostic discrepancy localization. The proposed residual learning adapts to varying inconsistent densities—such as focal lesions versus diffuse abnormalities—while cosine similarity fails to handle spatiotemporal heterogeneity. This is especially critical in chest imaging, where neighboring structures (e.g., pleural effusion vs. atelectasis) demand precise boundary distinction. On thoracic cases, FAMDR achieves 0.59 IoU compared to 0.49 for the cosine variant.

RQ3: Case Study

Output Comparison of Different Models. To better understand the performance differences among models, we present a detailed case study in Figure 3. The case involves a patient (ID: p18996191) with a documented right lower lobe lobectomy, where subsequent imaging reports suggested possible nodules in the anatomically absent region. CLIP, as a general-purpose vision-language model, focuses primarily on visual features and generates a report suggesting follow-up examination for nodules in the right lower chest, completely missing the anatomical impossibility due to prior surgery. MedKLIP, while leveraging medical domain knowledge, recognizes the surgical history but still generates an ambiguous report that fails to definitively address the contradiction between observed findings and surgical status. In contrast, FAMDR demonstrates superior performance by explicitly identifying the anatomical-historical

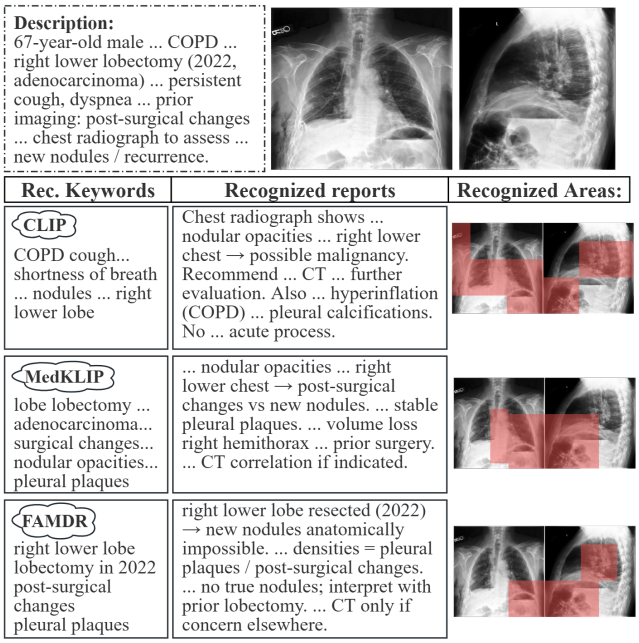

Figure 3: Comparison of CLIP, MedKLIP, and FAMDR outputs for a case involving COPD and prior right lower lobectomy highlights the importance of incorporating surgical history into image interpretation to prevent false positives and unnecessary follow-ups.

discrepancies. It not only recognizes that new nodules cannot exist in a surgically removed region but also provides a comprehensive analysis explaining that the observed "nodular" appearances likely represent post-surgical changes and pleural plaques. This case effectively illustrates FAMDR's unique capability to integrate temporal surgical history with anatomical knowledge, leading to more accurate and clinically relevant interpretations that prevent potential medical errors arising from historical-radiological discrepancies.

Cross-Modal Alignment Visualization. Figure 4 illustrates the feature space evolution through FAMDR's discrepancy-aware alignment. In the pre-alignment stage (left/middle), image patches (blue dots) and text tokens (red triangles) occupy distinct regions – imaging features cluster around [2,5] (e.g., pleural effusion patterns), while textual features aggregate near [5,2] (e.g., negation phrases). Post-alignment (right), both modalities converge into a shared semantic space with three clinically meaningful clusters: lung-related (green), heart-related (purple), and bone-related (orange). The directional arrows depict alignment trajectories, where inconsistent pairs (e.g., an effusion image with "no pleural fluid" text) are pulled toward cluster boundaries, preserving discriminative discrepancy features. This structured embedding space explains FAMDR's superior denoising in terms of F1 and IoU, as semantically coherent yet discrepancy-sensitive representations enhance detection robustness.

Clinical Semantic Preservation. The tight intra-cluster alignment (bone image-text distance reduced by 68% vs.

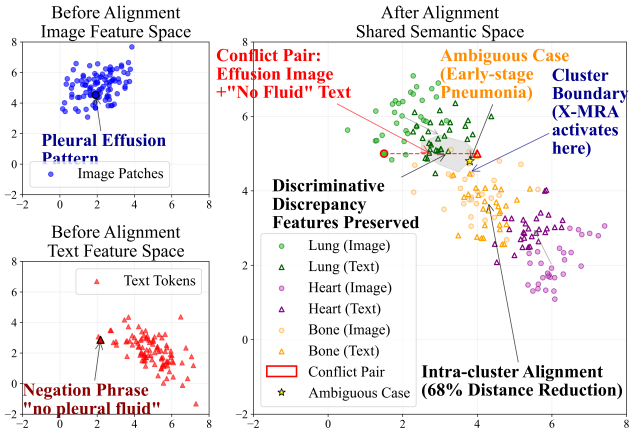

Figure 4: Alignment visualization. Pre-alignment (left) and shared semantic space (right) showing clinical clusters (lung, heart, bone), inconsistent pairs (red dashed), and ambiguous cases (yellow star) triggering retrieval-augmented reasoning.

pre-alignment) validates FAMDR's capacity to preserve clinical semantics while modeling discrepancies. For instance, bone abnormality descriptors ("fracture", "cortical disruption") co-locate with corresponding imaging biomarkers (e.g., trabecular discontinuity) in the orange cluster, enabling precise localization. Conversely, ambiguous cases (e.g., early-stage pneumonia) reside near cluster edges, where retrieval augmentation (Module X-MRA) activates to resolve uncertainty. This behavior balances semantic coherence and conflict separation with a proper λ setting, avoiding over-regularization.

RQ4: Modality Significance Study

Clinical diagnostics often involve reconciling multimodal inconsistencies. For example, a chest X-ray showing a lung opacity was initially interpreted as a "suspicious pulmonary mass," contradicting the patient's EHR, which documented prior breast implant surgery. Clinicians resolved the discrepancy by dynamically reweighting evidence—using the EHR to rule out malignancy, revisiting imaging features, and correctly identifying the opacity as a benign implant projection.

The modality significance study reveals crucial insights into the dynamic interplay between different information sources in chest X-ray diagnosis. As shown in Figure 5(a), we analyzed the contribution weights of two primary modalities - radiology assessment and EHR information - across three representative cases involving chest X-ray diagnostic challenges from the MIMIC-CXR dataset. The breast implant case demonstrates a particularly noteworthy distribution, with EHR information contributing 65% of the diagnostic weight and radiology assessment contributing 35%, reflecting the critical importance of historical medical records in preventing imaging misinterpretation.

Figure 5(b) illustrates the dynamic evolution of modality weights during the diagnostic process of the breast implant case, revealing a significant shift in the relative importance

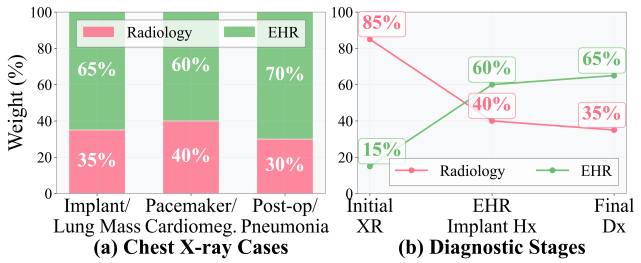

Figure 5: Modality contribution visualization in chest X-ray diagnosis. (a) Stacked bar chart shows modality contributions in three MIMIC-CXR diagnostic challenges, where historical EHRs dominate in resolving imaging misinterpretations. (b) In the breast implant case, diagnostic emphasis shifts from radiology (85%) to EHR-informed diagnosis—65% implant history and 35% refined imaging—underscoring the importance of integrating medical history with current imaging.

of different information sources. Initially, radiology assessment dominated the diagnostic weight (85%) due to its role in primary visualization. However, upon EHR review, the weight distribution dramatically shifted, with EHR information gaining prominence (65%) and radiology assessment's contribution decreasing to 35%. This weight redistribution reflects the decisive influence of historical medical documentation on imaging interpretation.

This analysis demonstrates that effective chest X-ray diagnosis requires a dynamic weighting mechanism that can adapt to emerging evidence and historical medical constraints. The EHR information's high contribution weight across all cases (ranging from 60-70%) emphasizes the fundamental role of historical medical documentation in resolving imaging misinterpretations. Moreover, the significant contribution of radiology assessment (30-40%) highlights the value of visual evidence in providing morphological details, suggesting that multimodal integration should prioritize historical medical constraints while maintaining flexibility in incorporating new imaging evidence.

RQ5: Parameter Sensitivity

Table 3 summarizes the impact of key hyperparameters on FAMDR's performance.

Alignment Loss Weight (λ). The contrastive-consistency tradeoff controlled by λ (Eq.1) exhibits a clear optimum at $\lambda = 0.5$, achieving peak F1=0.83. Lower values ($\lambda = 0.1$) under-regularize discrepancy features, allowing noisy discrepancies to distort cross-attention (F1=0.76). Conversely, higher weights ($\lambda = 1.0$) over-suppress genuine discrepancies, reducing F1 for subtle contradictions (e.g., misaligned anatomical references). This aligns with CAMAL's dual-stream design: balanced weighting preserves both modality-invariant semantics and discriminative conflict signals.

Retrieval Consistency Weight (β). Optimal retrieval grounding occurs at $\beta = 1.0$ (Eq.2), where external evidence complements original observations without dominating local context (F1=0.83). Over-reliance on retrieved knowledge ($\beta = 2.0$) introduces irrelevant references (e.g., conflating pleural effusion with atelectasis cases), dropping F1 to 0.77. Conversely, weak retrieval integration ($\beta = 0.5$) fails to resolve ambiguities in low-confidence scenarios (e.g., early-stage lesions), reducing precision. The balance validates Module X-MRA's gated fusion, which adaptively blends retrieved and local features based on discrepancy likelihood.

	Values	F1
λ	0.1, 0.3, 0.5, 0.7, 1.0	0.76, 0.80, 0.83, 0.79, 0.73
β	0.5, 1.0, 1.5, 2.0	0.78, 0.83, 0.81, 0.77
α	0.1, 0.2, 0.3, 0.4, 0.5	0.79, 0.81, 0.83, 0.80, 0.76

Table 3: Impact of key parameters on F1.

Localization Margin (α). The margin α (Eq.3) governs the separation between inconsistent and harmonious pairs. Optimal discrimination (F1 = 0.83) is achieved at $\alpha = 0.3$, where the model effectively balances sensitivity and specificity. Smaller margins (e.g., $\alpha = 0.1$) fail to suppress false positives in overlapping patterns such as peribronchial thickening versus fibrosis, while larger margins (e.g., $\alpha = 0.5$) impair the detection of fine-grained, spatially diffuse discrepancies like ground-glass opacities. These trends align with Module G-CLAM's adaptive thresholds, which calibrate conflict sensitivity based on anatomical complexity.

Conclusion

We propose FAMDR, a framework for resolving cross-modal inconsistencies between medical images and textual records. Unlike prior approaches that oversimplify contradictions or focus solely on alignment, FAMDR explicitly models discrepancies to enable context-aware, clinically coherent decisions. Evaluated on the extended MIMIC-CXR dataset, FAMDR outperforms the SOTA baselines in contradiction detection, localization, and uncertainty calibration. By supporting fine-grained, safety-aware interpretation, FAMDR enhances medical AI reliability and enables broader clinical integration.

While FAMDR shows strong performance on the extended MIMIC-CXR dataset, its generalizability to diverse clinical settings remains an open challenge. MIMIC-CXR, being a single-institution, U.S.-based dataset with standardized imaging and EHR protocols, may not capture the variability found in broader clinical practice. Other large-scale datasets (e.g., CheXpert, ChestX-ray14, PadChest) lack temporally aligned records or detailed inconsistency annotations, limiting their suitability for cross-modal contradiction detection. Moreover, differences in documentation style and terminology across institutions pose potential domain shift risks. Future work will explore domain adaptation, convention-aware retrieval, and synthetic perturbation to enhance robustness across clinical settings.

Acknowledgments

Supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (23XNL017).

References

Chen, Z.; Du, Y.; Hu, J.; Liu, Y.; Li, G.; Wan, X.; and Chang, T.-H. 2024. Mapping medical image-text to a joint space via masked modeling. *Medical Image Analysis*, 91: 103018.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021. Cross-modal Memory Networks for Radiology Report Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5904–5914.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1439–1449.

Dawidowicz, G.; Hirsch, E.; and Tal, A. 2023. Limitr: Leveraging local information for medical image-text representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21165–21173.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Balcan, M. F.; and Weinberger, K. Q., eds., *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, 1050–1059. New York, New York, USA: PMLR.

Gruber, S.; and Buettner, F. 2022. Better Uncertainty Calibration via Proper Scores for Classification and Beyond. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 8618–8632. Curran Associates, Inc.

Hou, W.; Cheng, Y.; Xu, K.; Li, W.; and Liu, J. 2023. RECAP: Towards Precise Radiology Report Generation via Dynamic Disease Progression Reasoning. In *2023 Findings of the Association for Computational Linguistics: EMNLP 2023*, 2134–2147. Association for Computational Linguistics (ACL).

Huang, K.; Altosaar, J.; and Ranganath, R. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Huang, Z.; Zhang, X.; and Zhang, S. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19809–19818.

Jin, H.; Che, H.; Lin, Y.; and Chen, H. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2607–2615.

Krishnan, R.; and Tickoo, O. 2020. Improving model calibration with accuracy versus uncertainty optimization. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 18237–18248. Curran Associates, Inc.

Lai, H.; Yao, Q.; Jiang, Z.; Wang, R.; He, Z.; Tao, X.; and Zhou, S. K. 2024. Carzero: Cross-attention alignment for radiology zero-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11137–11146.

Lemay, A.; Hoebel, K.; Bridge, C. P.; Befano, B.; De Sanjosé, S.; Egemen, D.; Rodriguez, A. C.; Schiffman, M.; Campbell, J. P.; and Kalpathy-Cramer, J. 2022. Improving the repeatability of deep learning models with Monte Carlo dropout. *npj Digital Medicine*, 5(1): 174.

Li, M.; Lin, B.; Chen, Z.; Lin, H.; Liang, X.; and Chang, X. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3334–3343.

Liu, B.; Wang, C.; Cao, T.; Jia, K.; and Huang, J. 2024. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7817–7826.

Mei, X.; Mao, R.; Cai, X.; Yang, L.; and Cambria, E. 2024. Medical Report Generation via Multimodal Spatio-Temporal Fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4699–4708.

Miura, Y.; Zhang, Y.; Tsai, E.; Langlotz, C.; and Jurafsky, D. 2021. Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5288–5304.

Nicolson, A.; Dowling, J.; Anderson, D.; and Koopman, B. 2024. Longitudinal data and a semantic similarity reward for chest X-ray report generation. *Informatics in Medicine Unlocked*, 50: 101585.

Sun, J.; Wei, D.; Xu, Z.; Lu, D.; Liu, H.; Wang, H.; Tsaftaris, S. A.; McDonagh, S.; Zheng, Y.; and Wang, L. 2024. Unlocking the Potential of Weakly Labeled Data: A Co-Evolutionary Learning Framework for Abnormality Detection and Report Generation. *IEEE Transactions on Medical Imaging*.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7433–7442.

Thawkar, O.; Shaker, A.; Mullappilly, S. S.; Cholakkal, H.; Anwer, R. M.; Khan, S.; Laaksonen, J.; and Khan, F. S. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Tulinský, L.; Ihnát, P.; Mitták, M.; Guňková, P.; and Zonča, P. 2016. Intrathoracic splenosis – lesson learned: a case report. *Journal of Cardiothoracic Surgery*, 11(1): 72.

Wang, F.; Zhou, Y.; Wang, S.; Vardhanabhuti, V.; and Yu, L. 2022. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35: 33536–33549.

Xu, D.; Zhu, H.; Huang, Y.; Jin, Z.; Ding, W.; Li, H.; and Ran, M. 2023. Vision-knowledge fusion model for multi-domain medical report generation. *Information Fusion*, 97: 101817.

Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.