Organ-Aware Routing Mixture-of-Retrieval Augmented Generation for Fetal Ultrasound Reporting

Bin Pu[1*], Siyu Wang[2*], Rongbin Li[2], Xinpeng Ding[3†], Lei Zhao[1], Chaoqi Chen[2], Shengli Li[4], Kenli Li[1†]

[1]Hunan University, Changsha, China
[2]Shenzhen University, Shenzhen, China
[3]The Hong Kong University of Science and Technology, HKSAR, China
[4]Shenzhen Maternity and Child Healthcare Hospital, Southern Medical University, Shenzhen, China
{pubin,zhaolei,lkl}@hnu.edu.cn, xdingaf@connect.ust.hk, cqchen1994@gmail.com

Abstract

Fetal ultrasound screening is a uniquely complex diagnostic task involving the simultaneous assessment of multiple fetal organs—each with its own anatomical and clinical context—within a single examination. Automating report generation in this setting is particularly challenging due to the **multiple-to-multiple misaligned correspondence** between organ-specific image sets and the multi-section report, *i.e.*, each report segment may relate to multiple images, but the exact alignments are unknown. In this paper, we introduce **FetusR**, the first large-scale dataset for multi-organ fetal ultrasound reporting, containing 15,594 real-world cases with rich organ-wise annotations. To address the alignment challenge, we propose ***Organ-Aware Routing Mixture-of-Retrieval Augmented Generation (ORM-RAG)*** inspired by the Mixture-of-Experts paradigm. ORM-RAG decomposes the complex alignment problem into multiple one-to-one sub-retrieval tasks. Specifically, ORM-RAG integrates (1) an organ-aware mixture-of-retrieval module that partitions the retrieval space into organ-specific corpora for independent retrieval, and (2) a dynamic routing mechanism that selectively aggregates high-confidence organ-specific reports while filtering uncertain ones. Extensive experiments demonstrate that ORM-RAG significantly outperforms state-of-the-art baselines across both textual similarity and clinical accuracy metrics. Our work opens a new direction for long-form, structured report generation in real-world, multi-organ medical imaging scenarios.

Introduction

Fetal ultrasound is the only safe and common imaging method for examining unborn babies (Quinn et al. 2023). Unlike adult or pediatric scans that focus on single organs, fetal screening evaluates multiple key organs (Ha et al. 2021; Khalil et al. 2024; Salomon et al. 2022)—brain, heart, face, abdomen, limbs, and placenta—in one exam, as shown in Fig. 1 (a). This thorough check is crucial for early detection of serious birth defects like heart malformations and brain anomalies. Consequently, reports are long, detailed, and organ-specific. Sonographers have a heavier workload than radiologists, needing to examine more structures and write longer reports, all while facing a global shortage of trained staff (Yan et al. 2025). These factors highlight the urgent need for automating fetal ultrasound report generation.

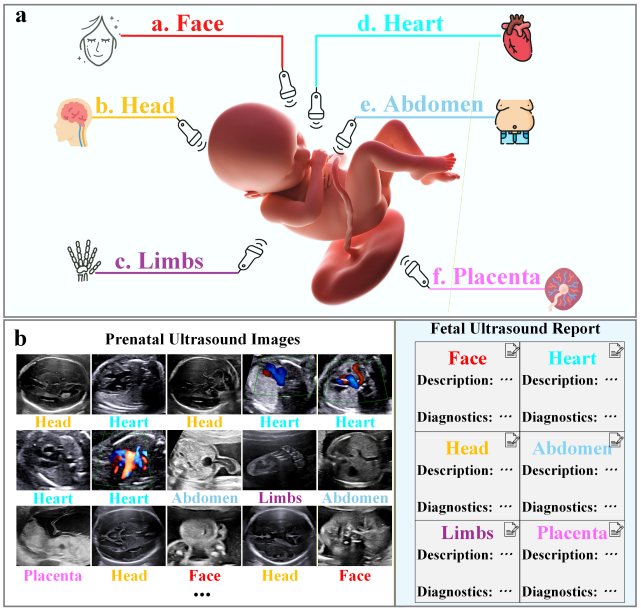

Figure 1: **Fetal multi-organ ultrasound report generation.** (a) The fetus undergoes a single examination covering multiple organs, each described in a report. (b) Each case contains imaging data of several organs, with multiple ultrasound images per organ. Notably, each report segment may relate to multiple images, but the alignments are unknown, posing a **multiple-to-multiple misalignment challenge.**

To advance the fetal ultrasound report generation, we introduce ***FetusR***, a first large-scale ultrasound dataset comprising 15, 594 fetal screening cases with corresponding diagnostic

*These authors contributed equally.
†Corresponding authors.

The appendices and supplementary materials are available at https://github.com/Si-yu-Wang/ORM-RAG.

reports. As shown in Fig. 1 (b), each case includes ultrasound images and corresponding diagnostic reports that cover six key organs: the face, head, heart, abdomen, limbs, and placenta. The reports provide a comprehensive foundation for developing automated multi-organ fetal ultrasound report generation, enabling research into long-form, multi-organ report generation in real-world settings.

Recent advances in large multimodal models (Li et al. 2024a; Liu et al. 2023; Chen et al. 2023; Lin et al. 2023a; Li et al. 2023b) have significantly boosted research on automated report generation. By leveraging vision-language pretraining and instruction tuning (Liu et al. 2023), these models show promising results in generating descriptive and coherent reports from radiological images (Li et al. 2023a; Wang et al. 2025; Liu et al. 2025). However, their performance in real-world clinical settings remains limited (Zhang et al. 2023). This is mainly due to the domain gap between general pretraining data and specialized medical data, as well as the difficulty of capturing fine-grained, localized abnormalities crucial for clinical decisions (Jeong et al. 2024).

Retrieval-Augmented Generation (RAG) (Endo et al. 2021; Jeong et al. 2024) offers a compelling way to enhance factual accuracy by grounding generation in retrieved, in-domain reports. While recent studies have applied RAG to medical report generation, they mostly focus on Chest X-rays or CTs (Jeong et al. 2024), which is a **one-to-one** setting, *i.e.*, a report only contains the information and diagnosis for a single organ. However, automated fetal ultrasound report generation is in a **multiple-to-multiple** nature, *i.e.*, each unified report corresponds to the information and diagnosis of multiple organs, and each organ includes multiple ultrasound images. Naively using previous RAG ways, *e.g.*, one-to-multiple or multiple-to-multiple retrieval strategies either suffer from misalignment or noisy (see details in Fig. 2 (a)-(b)). This calls for a new retrieval mechanism tailored to the structured nature of fetal ultrasound.

Motivated by the Mixture-of-Experts (MoE) paradigm (Masoudnia and Ebrahimpour 2014; Xue et al. 2024), which routes inputs to specialized experts for task-specific processing, we propose **ORM-RAG** (Organ-Aware Routing Mixture-of-Retrieval Augmented Generation), a novel RAG framework designed to decouple the *multiple-to-multiple* problem into the **mixture of multiple one-to-one sub-retrievals**, as shown in Fig. 2 (c). Our approach introduces two core components: **(a) Organ-aware mixture-of-retrieval** that decouples the retrieval space into multiple organ-specific corpora . This enables independent retrieval within each organ domain, effectively isolating retrieval from cross-organ interference. **(b) Dynamic routing** that assigns confidence scores to each retrieved report and selectively routes only high-confidence ones into the generation model . This filters out irrelevant or uncertain information, improving the diagnostic accuracy.

Our contributions are summarized as follows:

- We introduce ***FetusR***, the first large-scale dataset for fetal multi-organ ultrasound report generation with $15,594$ real clinical cases.
- We propose **ORM-RAG**, a novel organ-aware RAG framework that converts the multi-organ retrieval problem into a mixture of structured one-to-one retrievals.
- We demonstrate that ORM-RAG achieves state-of-the-art performance on both textual and diagnostic evaluation metrics, with comprehensive ablation studies validating its components.

Related Work

Automated Diagnosis of Fetal Ultrasound Diseases. Automated fetal ultrasound analysis has evolved along three main directions: (1) Knowledge-based methods, which leverage semantic modeling and graph structures (Lin et al. 2023b; Lu et al. 2024); (2) Hardware-integrated frameworks for clinical deployment (Carroll et al. 2024; Day et al. 2025); and (3) Architectural innovations, such as attention-based and state-space models (Lu et al. 2025; Gao et al. 2025; Nurmaini et al. 2025). For example, PAICS (Lin et al. 2023b) uses supervised learning to improve neurosonographic diagnosis, while SKGC (Lu et al. 2024) builds semantic knowledge graphs with contrastive learning. Mamba U-Net (Lu et al. 2025) introduces temporal modeling for fetal cardiac imaging, and (Nurmaini et al. 2025) enhances anomaly detection via self-attention and residual networks. Despite these advances, current methods face three limitations: (1) Sparse cross-organ semantics, with limited modeling of inter-organ dependencies; (2) Inconsistent feature representations due to anatomical variability across organs; (3) Weak cross-modal alignment, as most work focuses on image-level tasks rather than end-to-end report generation. *Overall, existing approaches mainly target single-organ segmentation or classification, lacking integrated, multi-organ diagnostic reporting—thereby failing to capture the full complexity of fetal assessment.*

Medical Report Generation. The evolution of medical report generation systems can be broadly categorized into three methodological phases in the existing literature: (1) Conventional deep learning approaches utilizing CNN/LSTM architectures (Anderson et al. 2018; Vinyals et al. 2015; Liu et al. 2019; Jing, Wang, and Xing 2020), (2) Transformer-based methods (Chen et al. 2020; Nooralahzadeh et al. 2021; Wang et al. 2023a; Huang, Zhang, and Zhang 2023; Alfarghaly et al. 2021; Yan et al. 2022; Wang et al. 2022), and (3) LLM-based frameworks (Li et al. 2025; Liu et al. 2024b, 2025; Wang et al. 2025; Zhang et al. 2024; Liu et al. 2024a; Li et al. 2024b; Lee et al. 2024; Wang et al. 2023a; Yan et al. 2023). While CNN/LSTM and Transformer-based approaches have demonstrated notable success in medical report generation, they no longer attract significant research attention. In contrast, LLMs have become a burgeoning research hotspot, offering enhanced performance in producing structured and medically precise reports. HC-LLM (Liu et al. 2025) introduces a novel longitudinal report generation framework that captures disease progression through time-shared and time-specific feature extraction with multimodal constraints. LLM-RG4 (Wang et al. 2025) introduces a flexible radiology report generation framework that overcomes fixed-task limitations through adaptive token fusion and instruction-following LLMs. LLaVA-Med (Liu et al. 2024a) designs an efficient

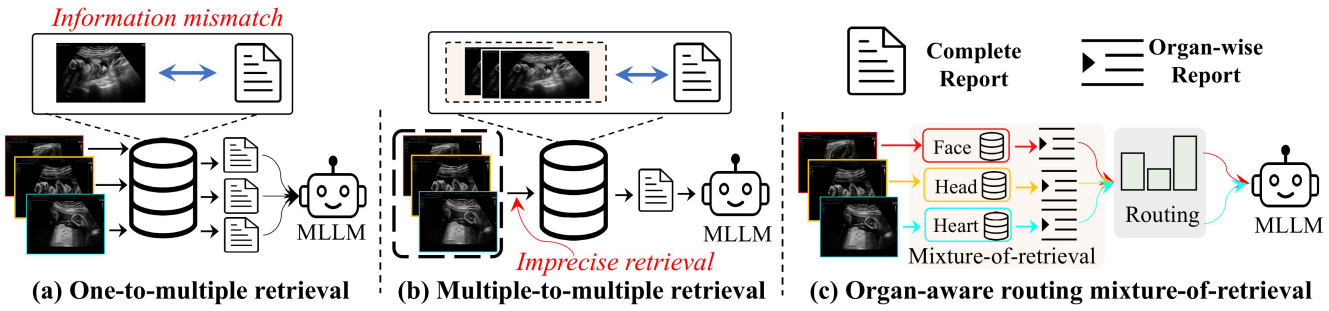

Figure 2: **Comparison of retrieval strategies. (a) Standard one-to-multiple retrieval** assigns each image to a full report, but in fetal ultrasound, a single image typically reflects only a small part of the report, leading to information mismatch. **(b) Multiple-to-multiple retrieval** aggregates the entire image set for the query, but abnormal findings are often limited to a few images. As a result, retrieval is dominated by normal-appearing regions, leading to imprecise retrieval and misleading context during generation. **(c) Organ-aware routing mixture-of-retrieval (multiple one-to-one)** decouples the retrieval space into organ-specific corpora and dynamically routes high-confidence organ-wise reports as the retrieval results. Note that for clarity, we only show three organs in this figure.

biomedical vision-language assistant trained through GPT-4 self-instruction and curriculum learning.

Recent studies (Jeong et al. 2024; Endo et al. 2021) have witnessed the emergence of retrieval-based report generation as a promising approach. For example, X-REM (Jeong et al. 2024) proposes a novel retrieval-based radiology report generation framework that leverages image-text matching scores to significantly improve both clinical accuracy and report relevance compared to conventional approaches. *However, previous work has been limited to generating reports for single-organ modalities (e.g., X-ray/CT) and has only addressed single-organ alignment, making these methods ill-suited for the complexities of multi-organ analysis.*

Preliminary

Retrieval-Augmented Generation (RAG) in Multimodal Report Generation

Multimodal large language models (MLLMs) have shown promise in medical report generation (Wang et al. 2023b) by processing visual inputs V and prompts T to produce structured reports $O = f(V, T)$. However, their factual accuracy remains limited due to the absence of domain-specific knowledge, making them susceptible to hallucinations—especially problematic in clinical applications (Zhang et al. 2023).

Retrieval-Augmented Generation (RAG) (Xia et al. 2024; Sun et al. 2024) addresses this by grounding generation on external knowledge retrieved from a domain-relevant corpus. A typical RAG pipeline includes: **1. Corpus Construction.** Relevant reports are sampled to construct $\mathcal{D} = \{X_i, Y_i\}_{i=1}^{N}$ of visual inputs and their corresponding reports to support similarity-based retrieval. **2. Query-Based Retrieval.** Given a query image V, a retrieval module $\mathcal{R}$ computes its embedding $q = E(V)$ and retrieves relevant reports

$$\begin{aligned} &\mathcal{Y} = \{Y_t \mid t \in \{\text{Top-K}_i(\text{sim}(q, e_i)) | i \in [1, N]\}\}, \\ &e_i = E(X_i) \end{aligned} \tag{1}$$

where $\text{sim}(q, e_i) = (q/|q|) \cdot (e_i/|e_i|)^\top$. $E(\cdot)$ is the vision encoder trained in a contrastive learning manner:

$$\mathcal{L}_{\text{con}} = -\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}\left(e_i, e_i\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(e_i, e_j\right)/\tau\right)}, \tag{2}$$

where τ is the temperature hyperemeters. **3. Generation.** The MLLM generates output based on the original inputs and the retrieved evidence: $O = f(V, T, \mathcal{Y})$.

Unique Challenges of Fetal Ultrasound Report Generation

Existing RAG frameworks have proven effective in radiology domains like chest X-ray (Sun et al. 2024), where each image corresponds to one report—a *one-to-one* setting, *i.e.*, $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$. However, as shown in Fig. 1, fetal ultrasound involves multiple images from various organ views contributing to a unified report consisting of several descriptions, constituting a *multiple-to-multiple misalignment*, *i.e.*, $\{\mathcal{X}_i, Y_i\}$, where $\mathcal{X}_i = \{X_{i,j}\}_{j=1}^{M_i}$, $X_{i,j}$ is the j-th image in $\mathcal{X}_i$. There are two straightforward ways to handle this, *i.e.*, *one-to-multiple* and *multiple-to-multiple*:

One-to-multiple retreieval. As shown in Fig. 2 (a), naively pairing each ultrasound image with the full report ($\{(X_{i,j}, Y_i)\}$). However, since each image typically maps only to a subset of findings, this way would lead to incomplete and noisy alignment.

Multiple-to-multiple retreieval. As shown in Fig. 2 (b), aggregating all images into a single embedding ($e_i = \frac{1}{M_i}\sum_j E(X_{i,j})$) then retrieving the full report. However, only a few images contain abnormalities, while most appear normal. Retrieving based on the entire image set risks being dominated by visually similar normal images, causing retrieval to prioritize irrelevant features over critical findings. This degrades retrieval precision and injects misleading context into generation.

Method

Motivated by the Mixture-of-Experts (MoE) paradigm (Masoudnia and Ebrahimpour 2014; Xue et al. 2024), which routes inputs to specialized experts for task-specific processing, we propose **ORM-RAG** (Organ-Aware Routing Mixture-of-Retrieval Augmented Generation), a novel RAG framework designed to decouple the *multiple-to-multiple* problem into the *mixture of multiple one-to-one sub-retrievals*, as shown in Fig. 2 (c).

Our approach introduces two core components: **(a) Organ-aware mixture-of-retrieval** that decouples the retrieval space into multiple organ-specific corpora . This enables independent retrieval within each organ domain, effectively isolating retrieval from cross-organ interference. **(b) Dynamic routing** that assigns confidence scores to each retrieved report and selectively routes only high-confidence ones into the generation model . This filters out irrelevant or uncertain information, improving the diagnostic accuracy.

Organ-Aware Mixture-of-Retrieval

In this section, we illustrate our organ-aware mixture-of-retrieval (O-MoR) approach to address the challenges, *e.g.*, information mismatch and imprecise retrieval. To achieve this, our O-MoR comprises two tailored designs: (a) an organ identification module to assign each image to a specific organ label; (b) a mixture-of-retrieval module to perform retrieval within each organ corpus individually.

Organ identification. Given a set of ultrasound images $\mathcal{V} = \{V_i\}_{i=1}^{M}$, we first classify each image into one of C anatomical organs using a trained classifier ϕ, i.e., $c_i = \phi(V_i)$. This enables routing each image to its corresponding retrieval space $\mathcal{D}^c$.

Corpus decoupling via organ-wise pairing. Similar to V_i, we also obtain the organ predictions for $X_i \in \mathcal{X}$, *i.e.*, $\phi(X_i)$. Then, we group the original image-report pair $(\mathcal{X}, Y)$ into C sub-pairs: $\{(\mathcal{X}^c, Y^c)\}_{c=1}^{C}$, where $\mathcal{X}^c$ contains images of organ c, and Y^c comprises report sentences related to c. This produces one-to-one pairs, avoiding information mismatch. Based on this, we can construct organ-specific corpora: $\mathcal{D}^c = \{(\mathcal{X}_i^c, Y_i^c)\}_{i=1}^{N^c}$, where N^c is the total number of paired image-reports in $\mathcal{D}^c$.

Retrieval optimization with abnorm-aware contrastive learning. In fetal ultrasound, the difference between normal and abnormal images can be visually subtle, and the standard contrastive learning in Eq.2 is hard to distinguish the normal and abnormal images accurately. To address this, we enhance the standard contrastive learning with abnormal-aware supervised learning. Specifically, given the embedding from the i, *i.e.*, $e_i^c = \frac{1}{M^c}\sum_{j=1}^{M^c} E(X_j)$, we have:

$$\mathcal{L}_{\text{ab-sup}} = -\sum_{i=1}^{N^c}\sum_{c=1}^{C} d_i^c \log h(e_i^c), \tag{3}$$

where i means the i-th image, $h(\cdot)$ is the classification head to predict e_i^c, $d_i \in \{0, 1\}$ is the label to indicate where the existing organ is normal 0 or abnormal (1). Finally, the optimization objective is defined as follows:

$$\mathcal{L} = \lambda * \mathcal{L}_{\text{organ-con}} + (1-\lambda) * \mathcal{L}_{\text{ab-sup}}. \tag{4}$$

Then, for images $\{V_i^c\}_{i=1}^{M^c}$ that are classified to the organ c, the retrieval process can be formulated as $O^c = \mathcal{R}^c(q^c, \mathcal{D}^c)$, where $q^c = \frac{1}{M^c}\sum_{i=1}^{M^c} E(V_i^c)$.

Dynamic Routing via Rank-Aware Consistency Estimation

In fetal report generation, directly feeding all retrieved examples into the LLM often introduces redundancy, noise, or even incorrect information, as evidenced in Table 5. To mitigate this, we propose a *dynamic routing mechanism* that filters retrieved results using a novel **rank-aware consistency estimation** module. This mechanism selectively routes only high-confidence retrievals into the LLM for report generation.

The key idea of the rank-aware consistency estimation is that if the top-K retrieved reports exhibit highly consistent diagnoses for the top-1 report O^c, the confidence of O^c is higher. Our approach combines two aspects of consistency: *global agreement* and *label trend continuity*.

Global Agreement. This metric captures whether the diagnostic labels $\{d_i\}_{i=1}^{K}$ converge around the top-1 label d_1. For each $i = 2, \ldots, K$, we define a binary match indicator:

$$m_i = \mathbb{I}(d_i = d_1), \tag{5}$$

and assign exponentially decaying weights to prioritize the top-ranked retrievals, presumed to be the most relevant:

$$w_i = \frac{\exp(-\beta i)}{\sum_{j=1}^{K}\exp(-\beta j)}. \tag{6}$$

We compute a smoothed, weighted match probability:

$$p_{\text{match}} = \frac{\sum_{i=2}^{K} m_i \cdot w_i + \gamma}{\sum_{i=2}^{K} w_i + 2\gamma}, \tag{7}$$

where γ is a Laplace smoothing constant that prevents overconfidence under sparse agreement. This ensures that p_{match} never equals exactly 0 or 1, maintaining numerical stability for the subsequent logarithm in Eq. 8. We then quantify the uncertainty of label agreement using binary entropy:

$$\alpha_{\text{Global}}^c = -p_{\text{match}}\log p_{\text{match}} - (1-p_{\text{match}})\log(1-p_{\text{match}}). \tag{8}$$

Lower entropy indicates stronger agreement and greater confidence in the top-1 result.

Label Trend Continuity. While global agreement captures overall support for d_1, it ignores sequential smoothness among the retrieved labels. To address this, we define a second metric that reflects how stably the diagnostic labels evolve across retrieval ranks. For $i = 2, \ldots, K$, we define a transition indicator:

$$t_i = \mathbb{I}(d_i = d_{i-1}), \tag{9}$$

and compute a trend continuity score that also considers agreement with d_1:

$$\alpha_{\text{Trend}}^c = \frac{\sum_{i=2}^{K} t_i \cdot w_i \cdot m_i}{\sum_{i=2}^{K} w_i}. \tag{10}$$

Method	*Textual Similarity*							*Diagnostic Accuracy*							
	B-1	B-2	B-3	B-4	M	R-L	CID	CA	PA	FA	LA	HA	AA	NM	Avg
***Without** Retrieval-Augmented Generation*															
LLaVA-1.5 (Liu et al. 2023)	0.67	0.65	0.63	0.61	0.37	0.71	2.96	0.0	35.5	0.0	0.0	0.0	0.0	69.4	37.6
LLaVA-Med (Li et al. 2023a)	0.71	0.69	0.67	0.65	0.39	0.75	3.35	2.6	40.4	0.0	0.0	0.0	0.0	72.2	39.7
R2GenGPT (Wang et al. 2023b)	0.80	0.75	0.72	0.70	–	0.74	3.55	25.4	41.7	0.0	0.0	0.0	0.0	66.7	38.9
MicarVLMoE (Izhar et al. 2025)	0.56	0.53	0.52	0.50	0.40	0.69	2.06	**85.9**	21.3	0.0	0.0	1.9	0.0	18.2	29.2
***With** Retrieval-Augmented Generation*															
CXR-RePaiR (Endo et al. 2021)	0.44	0.33	0.26	0.22	0.19	0.30	0.00	84.6	4.7	7.6	0.9	1.0	0.0	0.0	8.2
X-REM (Jeong et al. 2024)	0.33	0.31	0.30	0.29	0.25	0.56	1.84	0.5	11.2	0.0	7.6	1.0	0.0	95.7	49.5
BiomedCLIP (Zhang et al. 2023)	0.85	0.82	0.80	0.78	0.47	0.84	4.94	35.9	**69.2**	0.0	0.0	0.0	0.0	**99.5**	58.5
Ours	**0.86**	**0.83**	**0.81**	**0.79**	**0.47**	**0.85**	**4.99**	55.5	47.9	**31.9**	**34.9**	**25.0**	**6.4**	97.0	**65.3**

Table 1: **Comparison with state-of-the-art report generation methods on both textual similarity and diagnostic accuracy. All models are fine-tuned on our proposed *FetusR*, ensuring a fair comparison under identical training settings.** Left: BLEU (B-*), METEOR (M), ROUGE-L (R-L), CIDEr (CID). Right: CA (Cardiac), PA (Placental), FA (Facial), LA (Limb), HA (Head), AA (Abdominal) abnormalities, and NM (Normal). The best and second-best results are highlighted in **bold** and underline, respectively. **Our method consistently outperforms prior approaches across both textual and clinical metrics under the same setting.**

This score measures the temporal smoothness of the label sequence. *Unlike global agreement, it does not model uncertainty, so no entropy or smoothing is applied*. Instead, it directly rewards label transitions that are both consistent and aligned with the top-1 label.

Final Confidence Score. We combine the two metrics into a final confidence score using a sigmoid transformation:

$$\alpha^c = \frac{1}{1 + \exp\left(\sigma \cdot (\alpha^c_{\text{Global}} - \delta \cdot \alpha^c_{\text{Trend}})\right)}, \tag{11}$$

where σ controls the steepness of the curve, and δ balances the impact of trend continuity. High confidence emerges when the global entropy is low and the label trend is stable. In addition, this formulation allows trend instability to actively reduce the confidence derived from global agreement.

For all organ-wise retrieval candidate reports $\{O^c\}_{c=1}^{N^c}$, we compute a confidence score α^c for each using our rank-aware consistency estimation. We then select a subset of high-confidence retrievals to be fed into the LLM. Specifically, the final routed set $\mathcal{O}_{\text{final}}$ is defined as:

$$\mathcal{O}_{\text{final}} = \{O^c \mid \alpha^c > \tau,\ c = 1, \ldots, N^c\}, \tag{12}$$

where τ is a predefined confidence threshold. Only the elements in $\mathcal{O}_{\text{final}}$ are used as retrieval-augmented inputs to the LLM. When all retrieved examples yield confidence scores below the threshold τ,*i.e.*, $\mathcal{O}_{\text{final}} = \emptyset$, we discard the entire retrieval set to avoid introducing errors or hallucinated content and prepend a warning prompt to indicate that no reliable retrievals were available.

Experiment

Data Description

To fulfill the stipulated criteria for the automated generation of fetal multi-organ ultrasound reports, we collect a comprehensive dataset called ***FetusR*** from the Shenzhen Maternity and Child Healthcare Hospital. FetusR contains 15,594 confirmed prenatal cases, totaling 172,851 images, obtained from our partner hospital between January 2014 and March 2024. This study has been approved by the local hospital ethics committee (approval number: LLYJ2024-202-089). For details of *FetusR* dataset, see *Appendix B*.

Implementation Details

To get organ-specific sentences, we use a rule-based method to extract keywords and LLMs to refine the results. For our visual retriever, we use ViT-Base as the backbone pre-trained on ImageNet-21K. During training, with only its final two Transformer blocks being fine-tuned. We use the Adam optimizer with a learning rate of 1×10^{-5}, a batch size of 8, and train for 30 epochs. We index images in FAISS, a statistical bank for similarity searching, using 768-dimensional features obtained by mean-pooling patch tokens from the last ViT layer. Dynamic Routing is then applied to retain only confident, informative samples from the retrieved set. For the multimodal foundation model, we use InternVL as the backbone. All experiments were implemented using PyTorch and conducted on 4 NVIDIA A100 GPUs. More training details can be reviewed in *Appendix D*.

Comparison with state-of-the-art methods

To thoroughly evaluate the effectiveness of our proposed method, Table 1 shows the performance comparison with several state-of-the-art medical report generation approaches, including both RAG-based (CXR-RePaiR (Endo et al. 2021), X-REM (Jeong et al. 2024), and BiomedCLIP (Zhang et al. 2023)) and without RAG (LLaVA(Liu et al. 2023), LLaVA-Med (Li et al. 2023a), R2GenGPT(Wang et al. 2023b), and MicarVLMoE (Izhar et al. 2025)). **Note that all models here are fine-tuned on our proposed *FetusR*, ensuring a fair comparison under identical training settings.** Our evaluation considers two primary dimensions: **Textual Similarity** and **Diagnostic Accuracy**.

Textual similarity. This dimension measures how closely the generated text matches the reference reports. Table 1

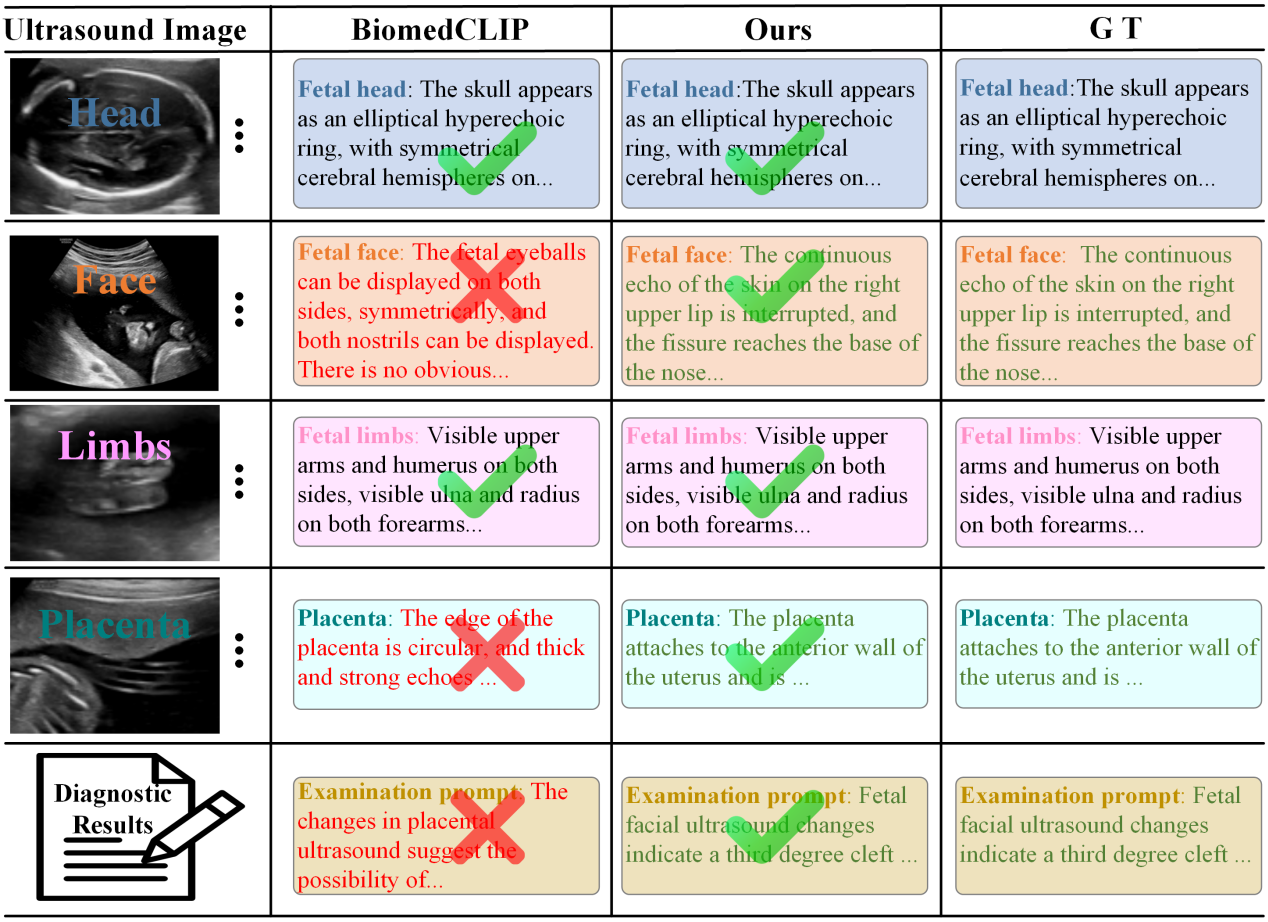

Figure 3: **Visualization comparison between different methods.** For clarity, here we only show a single ultrasound image for each organ and select three organs.

shows that our method achieves the best performance across all metrics. This high consistency with gold-standard reports shows the proposed approach can effectively generate accurate report descriptions, reduce physicians' workload, and holds strong potential for clinical application.

Diagnostic accuracy. Considering the high textual similarity does not necessarily guarantee clinical correctness or accurate diagnosis, we also introduce diagnostic accuracy, which assesses the model's ability to detect and classify medical abnormalities across multiple clinical categories.

Table 1 shows that SOTA models achieve strong textual similarity but suffer from poor diagnostic accuracy. This stem from (1) high visual similarity across multiple ultrasound images, making it hard to distinguish normal from abnormal, and (2) data imbalance, especially the scarcity of FA, LA, and AA abnormal samples, yielding near-zero accuracy for these categories. Our method overcomes these challenges by decoupling images by organs and applying organ-wise retrieval, boosting accuracy across all abnormality types.

Qualitative analysis. Fig. 3 compares report generations across multiple organs by different models. Our method produces coherent, organ-specific descriptions that align well with the ground truth (GT), demonstrating strong consistency and diagnostic accuracy. In contrast, BiomedCLIP (Zhang et al. 2023) shows limitations in capturing key abnormalities, often misdescribing malformations or omitting relevant details. These comparisons highlight our model's superior capability in handling multi-organ, long-form report generation. Additional visualizations are provided in *Appendix A*.

Ablation Study

We analyze the effect of different modules of our method. Here, we use InternVL2-1B (Chen et al. 2023) as our baseline model. In this section, we evaluate the performance from three aspects: retrieval accuracy-rank 1 (R@1) and rank 5 (R@5), textual similarity-BLEU-4 (B-4) and CIDEr (CID) and diagnostic accuracy-Abnormal accuracy (AB), normal accuracy (NM) and average accuracy (AVG).

Effect of each proposed module. Our method consists of two parts: mixture-of-retrieval (MoR) and dynamic routing (DR). Table 2 shows that MoR significantly boosts both textual similarity and diagnostic accuracy, demonstrating the effectiveness of organ-wise retrieval. Furthermore, using DR to remove noisy or irrelevant reports can improve the

Method	*Textual Similarity*							*Diagnostic Accuracy*							
	B-1	B-2	B-3	B-4	M	R-L	CID	CA	PA	FA	LA	HA	AA	NM	Avg
Baseline	0.794	0.762	0.739	0.721	0.454	0.793	2.293	48.3	73.0	0.00	4.70	1.90	0.00	46.5	33.9
w/ MoR	0.857	0.826	0.805	0.789	0.467	0.848	4.928	50.3	40.4	22.9	27.4	20.2	**8.5**	95.8	62.0
w/ MoR + DR	**0.859**	**0.829**	**0.808**	**0.793**	**0.470**	**0.853**	**4.992**	**55.5**	**47.9**	**31.9**	**34.9**	**25.0**	6.4	**97.0**	**65.3**

Table 2: **Ablation study of our method on both textual similarity and diagnostic accuracy.** Mor: Mixture-of-Retrieval (see in Section). DR: Dynamic Routing (see in Section). 'w/' means using specific modules. The best values are marked in **Bold**.

Method	*Retrieval*		*Textual*		*Diagnostic*		
	R@1	R@5	B-4	CID	AB	NM	AVG
Baseline	*Not*	*applicable*	0.721	2.293	21.3	46.5	33.9
O2M	0.569	0.883	0.741	3.848	21.8	82.8	52.3
M2M	0.545	0.857	0.707	3.560	20.9	66.7	43.8
M-O2O (Ours)	**0.764**	**0.917**	**0.793**	**4.992**	**33.6**	**97.0**	**65.3**

Table 3: **MoR ablation experiments.** 'O2M', 'M2M' and 'M-O2O' indicate the *one-to-multiple*, *multiple-to-multiple* and our *multiple one-to-one* setups, respectively.

Method	*Retrieval*		*Textual*		*Diagnostic*		
	R@1	R@5	B-4	CID	AB	NM	AVG
w/o	0.737	**0.928**	0.789	4.928	28.3	95.8	62.0
w/	**0.764**	0.917	**0.793**	**4.992**	**33.6**	**97.0**	**65.3**

Table 4: **The effect of abnorm-aware supervision.** 'w/o' and 'w/' mean without and with abnorm-aware supervision (Eq. 3) in the retriever optimization.

average diagnostic accuracy by 3.3%.

Ablation of Mixture-of-Retrieval Our MoR framework incorporates two designs tailored for fetal ultrasound. First, it decouples the many-to-many retrieval task into multiple organ-wise one-to-one retrievals. Second, it introduces an additional abnormality-aware supervision signal (Eq. 3) to enhance the training quality of the retriever. We conduct experiments to validate the effectiveness of these designs.

Effect of multiple one-to-one retrieval. Table 3 illustrates the comparison of different RAG methods shown in Fig. 2, *i.e.*, *one-to-multiple* (O2M), *multiple-to-multiple* (M2M) and our *multiple one-to-one* (M-O2O) respectively. Results show our M-O2O way achieves the best performance compared with other retrieval methods in all three metrics.

Effect of abnorm-aware supervision $\mathcal{L}_{\text{ab-sup}}$. In Section , we introduce the abnorm-aware supervised learning to enhance the learned retriever R^c to distinguish the abnorm and normal images. Table 4 shows that using $\mathcal{L}_{\text{ab-sup}}$ can improve the diagnostic accuracies across various organs.

Ablation of dynamic routing. We evaluated the confidence of retrieved reports from two aspects, *i.e.*, global agreement and label trend Continuity; see details in Section . Results from Table 5 show that global agreement improves diagnostic accuracy by avoiding errors or noisy retrieved reports in the final generation. Adding label trend continuity can further enhance the robustness of dynamic routing.

Method	*Retrieval*		*Textual*		*Diagnostic*		
	R@1	R@5	B-4	CID	AB	NM	AVG
w/o DR	0.737	0.928	0.789	4.928	28.3	95.8	62.0
α^c_{Global}	0.751	**0.918**	0.791	4.968	31.3	96.5	63.9
α^c_{Global}+α^c_{Trend}	**0.764**	0.917	**0.793**	**4.992**	**33.6**	**97.0**	**65.3**

Table 5: **Dynamic routing ablation experiments.** 'w/o DR' means no dynamic routing. Combining global agreement α^c_{Global} and label trend continuity α^c_{Trend} obtains the best.

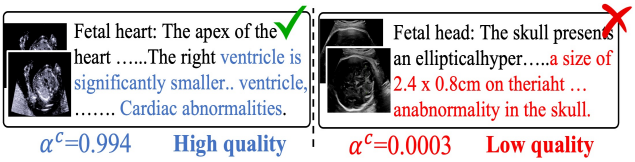

Figure 4: **Visualization of the relation of retrieved reports and confidence scores.** The retrieved report with a higher score (α^c) is more relevant to the input image's original report, and vice versa.

To further validate the relation of confidence score α^c and its retrieved report, we show two visualization examples in Fig. 4 (more detailed examples in *Appendix C*). The results show that high-quality retrieved reports typically have high confidence scores. In contrast, low-confidence reports include descriptions and diagnoses misaligned with the information in the input image (marked in red), indicating poor quality. This demonstrates the effectiveness of our dynamic routing.

Conclusion

This work pioneers automated fetal ultrasound report generation for multi-organ and multi-view analysis, addressing a critical gap in medical imaging where prior research focused solely on single-organ modalities like X-rays and CT scans. We introduce **FetusR**, a large-scale dataset spanning six fetal organ abnormalities, and propose ORM-RAG, a novel retrieval-augmented framework that tackles the challenge of aligning multiple-to-multiple image-text by decoupling retrieval into organ-specific tasks and dynamically routing high-confidence reports. integrating with a Multimodal Large Language Model (MLLM), our method generates detailed organ-wise reports and achieves state-of-the-art performance, outperforming existing MLLM-based methods. This advance enables scalable, structured reporting for complex prenatal ultrasound studies, with significant clinical deployment.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2025YFB3003705, in part by the National Natural Science Foundation of China under Grants 62227808, Grants 62506124, and in part by the Natural Science Foundation of Hunan Province under Grants 2025JJ60408.

References

Alfarghaly, O.; Khaled, R.; Elkorany, A.; Helal, M.; and Fahmy, A. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24: 100557.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Carroll, E.; Caracciolo, G.; Quintana, S. G.; Shelevytska, V.; Temko, A.; and Popovici, E. 2024. AI-Driven CHD Detection Using an Ultra-Low Power Embedded System. In *2024 35th Irish Signals and Systems Conference (ISSC)*, 1–6. IEEE.

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.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*.

Day, T. G.; Matthew, J.; Budd, S. F.; Farruggia, A.; Venturini, L.; Wright, R.; Jamshidi, B.; To, M.; Ling, H.; Lai, J.; et al. 2025. AI to Assist in the Fetal Anomaly Ultrasound Scan: A Randomized Controlled Trial. *NEJM AI*, 2(4): AIoa2400747.

Endo, M.; Krishnan, R.; Krishna, V.; Ng, A. Y.; and Rajpurkar, P. 2021. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, 209–219. PMLR.

Gao, Z.; Lin, Q.; Wen, H.; Pu, B.; Feng, M.; and Li, K. 2025. Incorporating Large Vision Model Distillation and Fuzzy Perception for Improving Disease Diagnosis. *IEEE Transactions on Fuzzy Systems*.

Ha, L. C.; Craig, A.; Grace, M. R.; Osmundson, S. S.; Taylor, E. W.; and Zuckerwise, L. C. 2021. Accuracy of estimated fetal weight assessment in fetuses with abdominal wall defects. *American journal of obstetrics & gynecology MFM*, 3(4): 100385.

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.

Izhar, A.; Japar, N.; Idris, N.; and Dang, T. 2025. MicarVLMoE: A Modern Gated Cross-Aligned Vision-Language Mixture of Experts Model for Medical Image Captioning and Report Generation. *arXiv preprint arXiv:2504.20343*.

Jeong, J.; Tian, K.; Li, A.; Hartung, S.; Adithan, S.; Behzadi, F.; Calle, J.; Osayande, D.; Pohlen, M.; and Rajpurkar, P. 2024. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Medical Imaging with Deep Learning*, 978–990. PMLR.

Jing, B.; Wang, Z.; and Xing, E. 2020. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. *arXiv preprint arXiv:2004.12274*.

Khalil, A.; Sotiriadis, A.; D'Antonio, F.; Da Silva Costa, F.; Odibo, A.; Prefumo, F.; Papageorghiou, A.; Salomon, L.; et al. 2024. ISUOG Practice Guidelines: performance of third-trimester obstetric ultrasound scan. *Ultrasound in Obstetrics & Gynecology*, 63(1): 131–147.

Lee, S.; Kim, W. J.; Chang, J.; and Ye, J. C. 2024. LLM-CXR: Instruction-Finetuned LLM for CXR Image Understanding and Generation. In *The Twelfth International Conference on Learning Representations*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Li, C.-Y.; Chang, K.-J.; Yang, C.-F.; Wu, H.-Y.; Chen, W.; Bansal, H.; Chen, L.; Yang, Y.-P.; Chen, Y.-C.; Chen, S.-P.; et al. 2025. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. *Nature Communications*, 16(1): 2258.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, Y.; Wang, Z.; Liu, Y.; Wang, L.; Liu, L.; and Zhou, L. 2024b. Kargen: Knowledge-enhanced automated radiology report generation using large language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 382–392. Springer.

Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; and Yuan, L. 2023a. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122*.

Lin, M.; Zhou, Q.; Lei, T.; Shang, N.; Zheng, Q.; He, X.; Wang, N.; and Xie, H. 2023b. Deep learning system improved detection efficacy of fetal intracranial malformations in a randomized controlled trial. *NPJ Digital Medicine*, 6(1): 191.

Liu, C.; Tian, Y.; Chen, W.; Song, Y.; and Zhang, Y. 2024a. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18635–18643.

Liu, G.; Hsu, T.-M. H.; McDermott, M.; Boag, W.; Weng, W.-H.; Szolovits, P.; and Ghassemi, M. 2019. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, 249–269. PMLR.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, T.; Wang, J.; Hu, Y.; Li, M.; Yi, J.; Chang, X.; Gao, J.; and Yin, B. 2025. HC-LLM: Historical-constrained large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5595–5603.

Liu, Y.; Wang, Z.; Li, Y.; Liang, X.; Liu, L.; Wang, L.; and Zhou, L. 2024b. MRScore: Evaluating Radiology Report Generation with LLM-based Reward System. *CoRR*.

Lu, Y.; Tan, G.; Pu, B.; Wang, H.; Liang, B.; Li, K.; and Rajapakse, J. C. 2024. SKGC: A General Semantic-level Knowledge Guided Classification Framework for Fetal Congenital Heart Disease. *IEEE Journal of Biomedical and Health Informatics*.

Lu, Y.; Tan, G.; Pu, B.; Yeung, P.-H.; Wang, H.; Li, S.; Rajapakse, J. C.; and Li, K. 2025. Optical Flow-Enhanced Mamba U-Net for Cardiac Phase Detection in Ultrasound Videos. *IEEE Transactions on Medical Imaging*.

Masoudnia, S.; and Ebrahimpour, R. 2014. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42: 275–293.

Nooralahzadeh, F.; Gonzalez, N. P.; Frauenfelder, T.; Fujimoto, K.; and Krauthammer, M. 2021. Progressive Transformer-Based Generation of Radiology Reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2824–2832.

Nurmaini, S.; Sapitri, A. I.; Roseno, M. T.; Rachmatullah, M. N.; Mirani, P.; Bernolian, N.; Darmawahyuni, A.; Tutuko, B.; Firdaus, F.; Islami, A.; et al. 2025. Computer-aided assessment for enlarged fetal heart with deep learning model. *iScience*, 28(5).

Quinn, L.; Tryposkiadis, K.; Deeks, J.; De Vet, H. C.; Mallett, S.; Mokkink, L. B.; Takwoingi, Y.; Taylor-Phillips, S.; and Sitch, A. 2023. Interobserver variability studies in diagnostic imaging: a methodological systematic review. *The British Journal of Radiology*, 96(1148): 20220972.

Salomon, L.; Alfirevic, Z.; Berghella, V.; Bilardo, C.; Chalouhi, G.; Costa, F. D. S.; Hernandez-Andrade, E.; Malinger, G.; Munoz, H.; Paladini, D.; et al. 2022. ISUOG Practice Guidelines (updated): performance of the routine mid-trimester fetal ultrasound scan. *Ultrasound in Obstetrics and Gynecology*, 59(6): 840–856.

Sun, L.; Zhao, J.; Han, M.; and Xiong, C. 2024. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. *arXiv preprint arXiv:2407.15268*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3156–3164.

Wang, Z.; Han, H.; Wang, L.; Li, X.; and Zhou, L. 2022. Automated radiographic report generation purely on transformer: A multicriteria supervised approach. *IEEE Transactions on Medical Imaging*, 41(10): 2803–2813.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023a. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11558–11567.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023b. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3): 100033.

Wang, Z.; Sun, Y.; Li, Z.; Yang, X.; Chen, F.; and Liao, H. 2025. Llm-rg4: Flexible and factual radiology report generation across diverse input contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8250–8258.

Xia, P.; Zhu, K.; Li, H.; Wang, T.; Shi, W.; Wang, S.; Zhang, L.; Zou, J.; and Yao, H. 2024. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Xue, F.; Zheng, Z.; Fu, Y.; Ni, J.; Zheng, Z.; Zhou, W.; and You, Y. 2024. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*.

Yan, B.; Liu, R.; Kuo, D.; Adithan, S.; Reis, E.; Kwak, S.; Venugopal, V.; O'Connell, C.; Saenz, A.; Rajpurkar, P.; et al. 2023. Style-Aware Radiology Report Generation with RadGraph and Few-Shot Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14676–14688.

Yan, B.; Pei, M.; Zhao, M.; Shan, C.; and Tian, Z. 2022. Prior guided transformer for accurate radiology reports generation. *IEEE Journal of Biomedical and Health Informatics*, 26(11): 5631–5640.

Yan, Y.; Wang, K.; Feng, B.; Yao, J.; Jiang, T.; Jin, Z.; Zheng, Y.; Zhou, Y.; Chen, C.; Sui, L.; et al. 2025. The use of large language models in detecting Chinese ultrasound report errors. *npj Digital Medicine*, 8(1): 66.

Zhang, L.; Liu, M.; Wang, L.; Zhang, Y.; Xu, X.; Pan, Z.; Feng, Y.; Zhao, J.; Zhang, L.; Yao, G.; et al. 2024. Constructing a large language model to generate impressions from findings in radiology reports. *Radiology*, 312(3): e240885.

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*.