*Q*Rad: Enhancing Radiology Report Generation by Captioning-to-VQA Reframing

Ying Jin[1,2] **Noel C. F. Codella**[1] **Yanbo Xu**[1] **Yu Gu**[1] **Mu Wei**[1]
Haoquan Fang[2] **Thomas Lin**[1] **Paul Vozila**[1] **Jenq-Neng Hwang**[2]

[1]Microsoft [2]University of Washington

{ying.jin, ncodella, yanboxu, aiden.gu}@microsoft.com
{muhsin.wei, tlin, paulvozila}@microsoft.com
{jinying, hqfang, hwang}@uw.edu

Abstract

Radiology Report Generation using AI has demonstrated significant potential in modern clinical workflows. However, existing approaches have limited clinical utility due to a lack of interactive capabilities and compromised factual reliability because linguistic variations are prevalent in the training data and lead to overfitting. We introduce *Q*Rad, a novel approach which reframes radiology report generation from image captioning to a self-directed visual question-answering (VQA) process. Specifically, we convert radiology reports into question-answer pairs and train our model to first generate a chain of questions and then respond with answers. The answers are concatenated to form the radiology report. Our approach offers three advantages: First, quality is considerably improved because sentence-level linguistic variations (such as the omission or ordering of medical topics) are removed from the answer generation's criterion, allowing the model to focus on factual accuracy rather than presentation style. Second, the model provides an intrinsic VQA capability that enables physicians to interact with the model for details that may have been omitted in the initial output. Third, *Q*Rad derives confidence scores from token probabilities through its ability to answer template questions about specific medical conditions, a capability unavailable in previous models, enabling Receiver Operating Characteristic (ROC) based evaluation to facilitate regulatory approvals. Experiments show that *Q*Rad outperforms state-of-the-art models with only 13% of their size, offering a promising path for clinical adoption and regulatory validation in real-world settings.

1 Introduction

Medical imaging plays a crucial role in healthcare diagnostics. However, the worldwide shortage of radiologists poses significant risks to patient care [16, 48, 5]. Automated radiology report generation using AI has emerged as a promising solution to this challenge, with the potential to reduce radiologist burden to only the most complex cases.

Despite recent advances in radiology report generation, significant gaps remain towards clinical adoption. First, current approaches, which typically follow an image captioning pipeline, struggle with the inherent linguistic uncertainties [55] in radiology reports. Unlike conventional image captioning, radiology reports are longer documents that require precise factual accuracy while exhibiting considerable sentence-level linguistic variation, such as whether a finding is mentioned

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

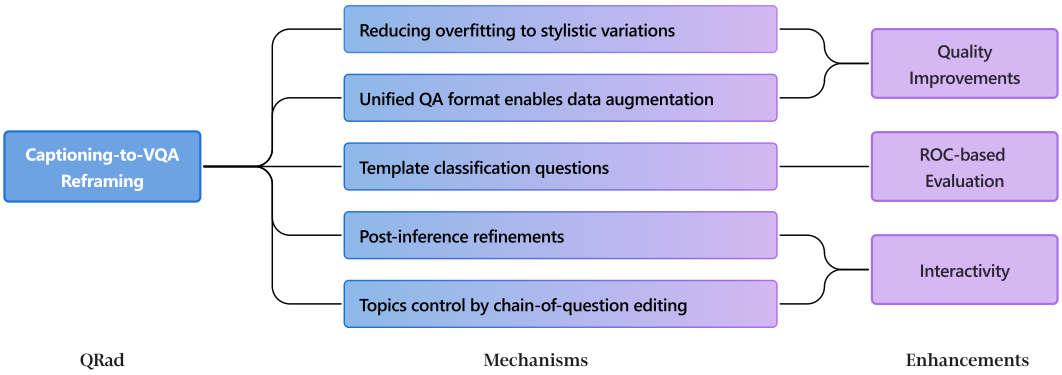

Figure 1: Overview of QRad's captioning-to-VQA reframing approach. This question-driven framework (the "Q" in QRad) enables five mechanisms that collectively enhance clinical utility across three dimensions: report quality, regulatory-required evaluations, and interactivity.

or omitted, and the order in which medical findings are presented. For example, if the ground truth has three sentences `[A, B, C]`, a prediction that reorders the same findings (e.g., `[C, A, B]`) is clinically correct but is unfairly penalized [21] by the language modeling loss because it requires exact token-by-token matches. Consequently, models tend to overfit such linguistic variations at the expense of factual accuracy. In report generation datasets such as MIMIC-CXR, each training sample contains one or multiple images and an associated text report. Image captioning datasets like COCO [35] provide multiple reference texts to capture the linguistic variances, however, this solution is not feasible in the collection of radiology report datasets. Furthermore, conventional approaches that follow a direct image-to-text pipeline [7, 56, 64, 10] offer no interactive mechanisms, preventing physicians from requesting additional information about specific concerns omitted in the initial output [47, 20].

Second, existing generative models lack the ability to produce continuous, numerical confidence scores for individual medical findings. For clinical utilization of software, FDA device authorization requires generating Receiver Operating Characteristic (ROC) curves and evaluating sensitivity and specificity across clinical applications with differing tolerances for false positives and false negatives [15]. For example, cancer screening prioritizes high sensitivity to avoid missed cases, while cohort discovery systems for clinical research require high specificity to accurately identify patients meeting strict inclusion criteria and reduce downstream noise. A model that produces confidence scores for requested disease classes can therefore facilitate regulatory approval, moving one step closer to real-world adoption.

To address these challenges, we introduce *Q*Rad, a novel approach that reframes radiology report generation as a self-directed visual question-answering (VQA) process. *Q*Rad operates in two steps: (1) Question Generation, which produces a chain of relevant clinical questions conditioned on the input radiograph, effectively planning the report's structure; (2) Answer Generation, which answers those questions by examining visual features. The answers are concatenated to form the final report. To facilitate training, we convert reference reports into QA pairs by segmenting each report into contiguous topical spans (answers), and GPT-4[1] generates a single question that captures each span's topic. This design offers two immediate benefits: it operationalizes Chain-of-Thought [61] via explicit planning-and-answering decomposition, and it provides an interactive capability, allowing physicians to request specific information beyond the initial report by editing or issuing follow-up questions — a feature unavailable in previous single-step models.

Formally, traditional approaches model report generation as $Y = f(I)$, where an image I directly maps to a report Y. Due to valid linguistic variations in the ground truth such as reordering of sentences, this formulation suffers from a one-to-many mapping from I to multiple valid Y, causing the learning process to overfit to surface phrasing at the expense of clinical accuracy. This limitation arises from the language modeling loss which treats each token equally, allowing the model to shortcut by producing a radiology report that achieves linguistic overlap with the ground truth on non-factual tokens while differing in key tokens that determine factual accuracy, such as presence/absence,

[1]We use a private, in-house deployment to satisfy data-usage requirements. The labeled data will be released.

severity, and location. *Q*Rad reframes the process as $Y = f_A(I, Q);, Q = f_Q(I)$. By providing the Answer Generator f_A with a question Q, we explicitly demand the model to state a diagnosis for the clinical topic. The ground truth for f_A is a single-sentence topical span (answer), reducing the space of linguistic variations and focusing on factual accuracy. The Question Generator f_Q captures linguistic variability—even when it produces questions that differ from the training data, these tend to be clinically valid variations that preserve diagnostic utility. In essence, we isolate linguistic variability in f_Q and concentrate factual supervision in f_A. Moreover, the VQA reframing allows us to augment the training data with additional image classification questions; in these cases, the ground truth is a single `Yes/No` token which further reduces linguistic variability and concentrates supervision on diagnostic accuracy.

Furthermore, typical regulatory processes (e.g., FDA approval) require ROC-based validation, which depends on class probabilities like those produced by perception models. *Q*Rad bridges this gap via *closed-vocabulary* VQA: for each predefined disease class, we pose a binary, template-based query (e.g., "`Is this image classified as [CLASS]? (yes/no)`"), extract the token logits for pre-defined answers `{Yes, No}`, and compute the softmax as class probabilities. In contrast, conventional report-generation models emit free-form sentences that may mention multiple diseases or omit a disease entirely, so token-level probabilities are not class-specific and cannot serve as per-class confidences to support ROC analysis. Meanwhile, image classifiers do not produce open-vocabulary reports that describe medical findings with flexibility. Our VQA reframing approach unifies both regimes, providing open-vocabulary narratives and closed-vocabulary class probabilities within a single backbone to support ROC/AUC analysis, offering a practical path toward regulatory clearance and real-world adoption.

In summary, we propose *Q*Rad, a captioning-to-VQA reframing approach that addresses key limitations in radiology report generation. As illustrated in Figure 1, our question-driven framework enables five core mechanisms that collectively enhance clinical utility across three critical dimensions: improving report quality by reducing overfitting to stylistic variations, enabling ROC-based evaluation through quantitative confidence scores, and providing interactivity enhancements via post-inference refinements and topic control. Experiments show that *Q*Rad outperforms state-of-the-art models [7, 56, 64, 10] while using less than 13% of their model size.

2 Related Work

2.1 Image Captioning

Image captioning aims to generate a sentence that describes a given image. The latest work benefits from large scale vision-language pre-training [8, 14, 58, 32]. Encoder-decoder architectures [34, 59, 41] provide a unified implementation for various vision-language tasks.

While many radiology report generation methods are based on image captioning [12, 57, 62, 66], there are key differences in the tasks including (1) radiology reports are much longer than generic image captions such as those in COCO Captions [35], and have multiple sentences covering different medical topics; (2) factual correctness is critical for radiology reports, which requires close examination of fine visual details; (3) image captioning datasets may provide multiple ground truths per image to capture linguistic variations, however, this is not available in radiology report datasets.

2.2 Radiology Report Generation

Chest X-ray radiology reports lack a standardized order for presenting medical findings [4]. For instance, the inside-out order [53] and the ABCDE order (each letter represents an anatomical region) [38] are two approaches from clinical guidelines. Additionally, medical conditions can be omitted from the report [23]. These valid linguistic variations lead to Loss-Metric mismatch problems, creating challenges for both training and evaluation [18, 65, 17]. Existing state-of-the-art methods use the original radiology reports as supervision and train the models in an image captioning setup, differing primarily in datasets, architectures, and pretraining/fine-tuning regimes. For example, [3] explores mimicking clinical setups and other methods leverage pre-training and fine-tuning techniques [64, 40, 4].

In connection to our study, previous studies also demonstrated two-step approaches [45, 36, 63, 27]. Specifically, Liu et al. [36] adopts a hierarchical framework that predicts sentence-level topics as

the first step. However, their topic generation module is not supervised with any labels, leaving uncertainty in their actual meaning. Nooralahzadeh et al. [45] first generates high-level context sentences and then refines them into the reports. The first step is trained to generate medical keywords per sentence that are extracted using a text processing model. We differ from them in the supervision of the first step. Yan et al. [63] replaces full reports with serialized RadGraph representations (entities and attributes) as supervision, thereby filtering out non-semantic words. In contrast, *Q*Rad addresses sentence-level style variations such as omission and reordering of findings, which RadGraph-based supervision still encodes. Jin et al. [27] generates questions and their answers, and use the QA pairs to prompt report generation for better grounding. The QA data are from external datasets such as [19]. In our method, we focus on the idea of task reframing from image captioning to a chain of VQA, where the answers are concatenated to be the captioning output. Therefore, our generated questions are ordered, matching the report sentence-by-sentence.

3 Method: Reframing Long Text Generation to Chain-of-VQA

Conventional approaches to long text generation from visual inputs frame the task as direct image-to-text mapping i.e., image captioning. As valid linguistic variations are prevalent in radiology reports, amplified by their length, factual accuracy is hindered when the model attempts to overfit the linguistic variations. We propose a general approach that reframes long text generation into a self-directed visual question-answering process, where the self-generated questions serve as an explicit plan akin to chain-of-thought [61] models.

The proposed Captioning-to-VQA reframing method is applicable to different model architectures. In our experiments, it effectively elevates the performance of a small model to match those 10X larger. In this section, we demonstrate our method with MIMIC-CXR [30], one of the largest radiology report dataset that is publically available.

3.1 Dataset Preparation

*Q*Rad leverages two types of question-answering datasets, including a report generation QA dataset derived from the image-report dataset, and an image classification QA dataset converted from image-class labels. Compared to training on the original full reports, the first dataset has reduced linguistic variations at the sentence-level (such as the omission and ordering of sentences), while the second dataset, being closed-vocabulary (the answers being `{Yes, No}`), further reduces linguistic variations to the phrase level.

3.1.1 Report Generation Question-Answer Pairs

- `[Q1] "What type of view is used in the chest X-ray?"`
  `[A1] "Single AP view of the chest provided."`
- `[Q2] "Are there any support devices visible?"`
  `[A2] "An endotracheal tube ends 2.0 cm above the Carina. A transesophageal tube courses below the level of the diaphragm, however the tip cannot be visualized."`
- `[Q3] "What is the condition of the lung volumes and clarity?"`
  `[A3] "Lung volumes are low, however grossly clear."`
- `[Q4] "Is there any atelectasis?"`
  `[A4] "Bibasilar atelectasis is moderately increased."`
- `[Q5] "Are there signs of pleural effusion or pneumothorax?"`
  `[A5] "No pleural effusion or pneumothorax."`

Figure 2: Example of the converted report generation QA dataset. We show the first five sentences from a radiology report, where Q_i and A_i are the i^{th} question and answer, respectively.

To convert such datasets to VQA format, we ask GPT-4 [46] to split the reports into sentence groups. Consecutive sentences in a report covering the same topic are treated as a cohesive unit. Then, we use each sentence group as an answer, and compose a corresponding question with GPT-4 [2]. Therefore, we convert an image captioning training sample to a sequence of VQA samples (see Figure 2 for an example). When generating the questions, we instruct the questions to be precise enough to indicate the topics while not being too specific to leak the answer. GPT-4 only operates in the offline data preparation step, and our model is trained to produce the questions and their answers in test time.

For the MIMIC-CXR [28] dataset, we generated a total of 818,867 question-answer pairs across all radiology studies. There are 110,959 unique questions (based on string matches, not semantic similarity). 91.3% of the reports have no more than 5 sentences, and 99.4% of the reports have no more than 8 sentences. Typical answers contain only one sentence.

3.1.2 Image Classification Question-Answer Pairs

One benefit of our Captioning-to-VQA reframing is the ability to unify different supervisions into the same VQA format, allowing our model to seamlessly learn from both kinds of annotations to achieve superior performance. Here we augment image-report data with image-class labels. Specifically, in addition to the report generation QA pairs, we convert image class labels (obtained from VisualCheXbert [25]) into the VQA format. This integration not only enhances our model's image understanding capabilities but also improves its ability to handle diverse input questions while providing a natural mechanism for confidence score extraction.

- [Q1] "Is this image classified as cardiomegaly? (yes/no)"
  [A1] "Yes"
- [Q2] "Does this chest X-ray demonstrate edema? (yes/no)"
  [A2] "No"
- [Q3] "Is pleural effusion evident in this chest X-ray? (yes/no)"
  [A3] "Yes"
- [Q4] "Does this radiograph indicate pneumothorax? (yes/no)"
  [A4] "No"
- [Q5] "Does this chest X-ray reveal support devices? (yes/no)"
  [A5] "No"

Figure 3: Example of question-answer pairs converted from image classification labels. The questions are formulated using question templates and pre-defined class names, with a "(yes/no)" suffix that distinguishes them from report generation QA pairs and indicates a single-token binary answer is expected.

As shown in Figure 3, classification labels are transformed to closed-vocabulary QA pairs using the 14 categories from CheXpert [23]. Questions are constructed by randomly sampling from a template pool. The closed-vocabulary nature of these QA pairs focuses on training the model's image classification capabilities like an image classifier. When training on such datasets, the model gets no reward for writing a full sentence that has token-wise overlap with the ground truth sentence but is factually incorrect.

3.2 VQA Pipeline and Model Architecture

*Q*Rad decomposes the traditional image-to-text generation task from $Y = f(I)$ into two distinct components: a Question Generation Module $Q = f_Q(I)$ and an Answer Generation Module $Y = f_A(I, Q)$, where I, Q, Y denote the input image, questions, and answers (sentences in the report), respectively. Both modules utilize identical transformer architectures: a MI2-based [11] visual backbone and a tiny text decoder of six transformer layers.

[2]We use a private, in-house deployment of GPT-4 to ensure compliance with the dataset usage requirements.

3.2.1 Question Generation Module

The question generation module conducts sequence generation autoregressively with reference to the previously generated questions. Concretely, it generates m output tokens $Q = (q_1, q_2, \ldots, q_m)$ by modeling Equation 1:

$$P(Q \mid X) = \prod_{i=0}^{m+1} P(q_i \mid X, q_0, q_1, \ldots, q_{i-1}), \tag{1}$$

The ground truth Q is the concatenated questions. When providing inputs to the Answer Generator, we split Q by the question mark ("?") to obtain individual questions.

3.2.2 Answer Generation Module

The Answer Generation learns to generate a sentence of n tokens $Y_i = (y_{i_1}, y_{i_2}, \ldots, y_{i_n})$ conditioned on the image and a question Q_i. Mathematically, the module models the following:

$$P(Y_i \mid X, Q_i) = \prod_{j=0}^{n+1} P(y_{i_j} \mid X, Q_i, y_{i_0}, y_{i_1}, \ldots, y_{i_{j-1}}), \tag{2}$$

where Y_i denotes the i^{th} answer corresponding to question Q_i. By iterating Q_i through all questions, the Answer Generator generates n sentences $Y = (Y_1, Y_2, \ldots, Y_n)$ and composes the whole radiology report. In the interactive VQA mode, Q_i is replaced by the tokenized user-entered question.

3.2.3 Training Recipe

Training Stages. We use MedImageInsight (MI2) [11] as the vision encoder, a 0.36B-parameter model trained on medical images. For text decoders, we use a six-layer, randomly initialized transformer text decoder of 0.07B parameters. The encoder and decoder are connected via a linear projection layer. The total model size is 0.9B, around 13% of current state-of-the-art models that are based on 7B parameter models. We pre-train the encoder and decoder (the encoder is frozen) on CXR-697K, an image-text pre-training dataset used in existing work [7]. Then, we duplicate the model and fine-tune for question and answer generation tasks, where the full model is made trainable.

Mixture of VQA Data. As discussed in subsection 3.1, our training data contain both report generation QA data and image classification QA to improve model performance. The data mixture ratio is discussed in Table 5.

Prompt Templates. Following existing studies, we use a short instruction which includes the Indication section when generating the questions and corresponding answers. The Indication section specifies the goal of the radiology study. We use the ground truth questions as input when training the Answer Generator.

Attention Masks for Training Efficiency. After converting the training data from image-report to image-QA pairs, the number of training samples increases by the number of sentences per report, which significantly increases training cost. To improve training efficiency, we concatenate all QA pairs for the same image and construct attention masks to control context visibility, thereby enabling us to run forward in one pass.

3.3 Numerical Class Probability Extraction

*Q*Rad enables producing numerical class probabilities for medical findings, a capability absent in conventional report generation models. The method is agnostic to model architectures, however, the proposed captioning-to-VQA reframing is a prerequisite to dedicate classification to a single `[yes]/[no]` token. In contrast, prior approaches represent binary classifications with free-text sentences that can span multiple tokens and display a high degree of stylistic variance, which makes extraction of confidence technically challenging.

To extract these class probabilities, we leverage our VQA architecture by sending template classification questions to the model and request binary "yes" or "no" answers. We deliberately designed

these responses to be single-token outputs, allowing us to extract clean probabilities directly from the model's output distribution. The confidence score for each class is computed using:

$$P(C_i = 1) = \frac{e^{x_{\mathbf{yes}}}}{e^{x_{\mathbf{yes}}} + e^{x_{\mathbf{no}}}}, \tag{3}$$

where $P(C_i = 1)$ represents the probability of the i^{th} class, calculated from the softmax over logits $x_{\mathbf{yes}}$ and $x_{\mathbf{no}}$, the logits of `[yes]` and `[no]` being generated as the next token. This approach effectively transforms text generation over a binary vocabulary into a proxy for image classification, while sharing the same model weights with the report generation mode.

The resulting class probabilities accurately reflect the model's intrinsic image understanding capabilities, though generated separately with the reports. [31] validates the method of using $P(\texttt{True})$ to assess a language model's intrinsic capability. We make one step forward to use the softmax concerning no other tokens but only `[yes]` and `[no]`. A ROC evaluation based on these class probabilities is provided in Appendix C.

3.4 Experiments and Ablation Studies

We conduct experiments on MIMIC-CXR [29, 30], the largest radiology report generation dataset. It has 227,835 image-report pairs. We use only the frontal view radiograph from each training sample. Following recent studies [7, 22], we use the IU X-ray dataset [13] as a fully held-out evaluation set. All 3198 frontal-view X-rays are used as the testing split unseen during training.

3.4.1 Radiology Report Generation

In Table 4 and Table 2, we evaluate our method on the official testing split of MIMIC-CXR using both lexical metrics and clinical efficacy (CE) metrics following existing work. With our VQA-based pipeline and the training recipe, we outperform existing state of the arts with only 13% the model size while providing medical VQA as an additional feature (Table 4); Results on the IU X-ray dataset is available in Appendix D. We include model training details and introduction to evaluation metrics in Appendix A.

The ROC-based evaluation per class is provided in Appendix C. Qualitative examples of generated questions and answers are included in Appendix B.

3.4.2 Ablation Study and Hyper-parameters

Effectiveness of each component: In Table 3, we conduct an ablation study on a MI2-based small model with 0.9B parameters. We first reframe the report generation task as a VQA process ("Caption-to-VQA"), and then augment the training data with image classification QA pairs ("Classification QA"). The table shows that each method brings consistent performance gains. The largest improvements are observed on Clinical Efficacy metrics (CheXbert, RadGraph), which reflect factual accuracy in the medical domain.

Appendix E - Table 5 compares implementation details of *Q*Rad across three dimensions, including the data mixture ratio, the source of pseudo-labels for the classification QA data and whether previous QA pairs are provided as input context. The conclusions drawn from comparing experiments (a) to (e) are:

- Performance is robust to data mixture ratios between 20% and 40%; From comparison: (a) vs. (b)
- Using P+U as the positive label, which aligns with the "CheXbert: uncertain as positive" evaluation, leads to consistent performance gains across metrics. This is likely due to uncertain labels being corresponded to diseases mentioned in prior studies but ambiguously stated in current reports; From comparison: (e) vs. (b), (c)
- Providing previous QA pairs as context improves performance; From comparison: (d) vs. (b), (c), (e)

Benefits of the Question Generator: Appendix E - Table 6 demonstrates the importance of using a learned Question Generator over fixed template questions. The key challenge in medical report

Table 1: Report Generation Performance on MIMIC-CXR

Model	CheXbert ("uncertain" as *negative*)				CheXbert ("uncertain" as *positive*)				RadGraph	BLEU		ROUGE
	Micro-avg		Macro-avg		Micro-avg		Macro-avg					
	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	ER	(1)	(4)	(L)
Multi-Image w/ Prior Study												
MAIRA-2 [3] [a]	**58.1**	**59.1**	**41.6**	**50.4**	**58.5**	**61.3**	**45.8**	**54.6**	**39.6**	**46.0**	**23.1**	**38.4**
Single Image, Model size $\geq$ 7B												
LLaVA-Rad [7] [F]	**57.3**	57.4	39.5	47.7	**57.3**	**60.2**	**44.0**	**53.3**	29.4	38.1	15.4	30.6
Med-Gemini [64] [F]	-	-	-	-	-	-	-	-	-	-	20.5	28.3
VILA-M3 40B [40]	-	-	-	-	-	-	-	-	-	-	**21.6**	**32.2**
Med-PaLM M [56]	53.6	**57.9**	**39.8**	**51.6**	-	-	-	-	-	32.3	11.3	27.3
MAIRA-1 [22] [F]	55.7	56.0	38.6	47.7	55.3	58.8	42.3	51.7	**29.6**	**39.2**	14.2	28.9
GPT-4V	35.5	25.8	20.4	19.6	35.6	33.3	25.3	29.6	13.2	16.4	1.9	13.2
CheXagent [10]	39.3	41.2	24.7	34.5	39.4	42.1	27.3	35.8	20.5	16.9	4.7	21.5
LLaVA-Med [33] [F]	27.2	22.0	15.5	16.6	27.3	24.4	18.7	20.5	6.5	22.2	1.0	13.3
LLaVA [37] [F]	22.9	23.4	15.4	17.5	23.7	26.9	17.0	20.3	4.5	21.0	1.3	13.8
Single Image, Model size = 4B												
Baseline [b]	54.3	55.2	36.9	46.6	54.1	57.4	40.4	50.5	**31.1**	40.1	**17.8**	**32.7**
***Q*Rad** (ours, 4B) [b]	**57.6**	**59.0**	**40.8**	**51.0**	**57.1**	**61.4**	**44.3**	**54.4**	**31.1**	**40.6**	17.5	32.5
Single Image, Model size <1B												
PromptMRG [26]	-	-	-	-	-	-	-	-	-	-	11.2	26.8
Flamingo [2]	-	-	-	-	51.9	56.5	-	-	-	-	10.1	29.7
CvT2Dist. [42]	44.2	-	30.7	-	-	-	-	-	-	39.3	12.7	28.6
$\mathcal{M}^2$ trans [39]	-	-	-	-	-	56.7	-	-	-	-	11.4	-
RGRG [54]	-	-	-	-	-	54.7	-	-	-	37.3	12.6	26.4
R2Gen [9]	-	-	-	-	22.8	34.6	-	-	-	35.3	10.3	27.7
TieNet [60]	-	-	-	-	-	27.1	-	-	-	-	8.1	-
MI2 [11]	56.3	57.9	38.4	49.3	55.7	59.3	43.2	52.1	28.5	37.3	15.3	31.7
***Q*Rad** [F] (ours, 0.9B)	**58.4**	**59.5**	**41.5**	**51.8**	**57.9**	**62.2**	**45.1**	**55.2**	**31.5**	**40.0**	**16.9**	**32.5**

[*] We highlight the best score under each model size category.
[F] The testing set includes only frontal-view images.
[a] The MAIRA-2 benchmark is redesigned to reflect clinical scenarios by combining multiple images from the same case into a single instance. Therefore, direct comparisons to other approaches cannot be made.
[b] The 4B models use BiomedCLIP [67] as the vision encoder and Phi-3-mini [1] as the text decoder.

generation is the vast and complex space of possible medical conditions that can appear in an image. It is infeasible to enumerate all potential diseases as a predefined set of template questions. Moreover, even if such an exhaustive list existed, requiring the model to answer questions about every possible condition would be computationally prohibitive and inefficient. Our Question Generator addresses this by dynamically predicting relevant questions based on the input image, focusing only on conditions likely to be present.

Quality of generated questions: We observe that when given oracle questions that clearly specify each sentence's topic, the model shows substantial performance gains (Appendix E - Table 6). This demonstrates that stylistic variations (omissions, reordering) in the training data create noisy supervision signals, causing prior models to memorize surface patterns rather than learn medical content. Our model's strong performance with oracle questions proves it generates factually accurate answers. The differences between oracle and predicted questions represent legitimate stylistic choices rather than errors—these variations are natural in clinical practice.

Table 2: Performance on the ReXrank Benchmark

Model	1/RadCliQ-v1	BLEU	BertScore	SembScore	RadGraph	RaTEScore	GREEN
UniRG-CXR[3]	**1.217**	0.248	**0.493**	**0.487**	**0.265**	**0.596**	0.352
***Q*Rad-0.9B**, ours	1.143	**0.264**	0.482	0.479	0.243	**0.596**	**0.362**
MedVersa [69]	1.103	0.209	0.448	0.466	0.273	0.550	0.374
Libra [68]	0.898	0.232	0.402	0.403	0.218	0.523	0.356
RadPhi3.5Vision [50]	0.888	0.223	0.386	0.431	0.207	0.534	0.294
CXRMate-ED [44]	0.872	0.208	0.383	0.396	0.223	0.531	0.327
CXRMate-RRG24 [43]	0.870	0.198	0.367	0.423	0.220	0.521	0.338
CheXpertPlus-CheX [6]	0.805	0.142	0.367	0.379	0.181	0.490	0.305
DD-LLava-X[3]	0.801	0.154	0.348	0.402	0.182	0.505	0.301
RaDialog [49]	0.799	0.127	0.363	0.387	0.172	0.485	0.273
CheXpertPlus-MIMIC [6]	0.788	0.145	0.361	0.375	0.170	0.485	0.311
RGRG [54]	0.755	0.130	0.348	0.344	0.168	0.491	0.273
MedGemma [51]	0.744	0.165	0.346	0.339	0.159	0.549	0.293
CheXagent [10]	0.741	0.113	0.346	0.347	0.148	0.474	0.257
MoERad [3]	0.726	0.163	0.341	0.334	0.143	0.465	0.240
Cvt2distilgpt2 [42]	0.719	0.126	0.331	0.329	0.149	0.432	0.268

[1] Results shown are for the Findings Generation task on the MIMIC-CXR dataset.
[2] Models are ranked by 1/RadCliQ-v1 (higher is better for all metrics).
[3] UniRG-CXR, DD-LLava-X and MoERad from the leaderboard have no publications available yet.

Table 3: Ablation Study on MIMIC-CXR

Model	CheXbert								RadGraph	BLEU		ROUGE
	("uncertain" as *negative*)				("uncertain" as *positive*)							
	Micro-avg		Macro-avg		Micro-avg		Macro-avg					
	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	ER	(1)	(4)	(L)
Baseline (MI2)												
median	56.2	57.8	38.3	49.2	55.7	59.2	42.1	52.0	31.1	37.3	15.3	31.7
ci_l	55.1	56.2	36.7	47.1	54.7	57.8	40.6	50.6	30.5	36.8	14.9	31.2
ci_h	57.3	59.4	40.0	51.3	56.7	60.7	43.5	53.7	31.8	37.8	15.7	32.1
Baseline + Captioning-to-VQA												
median	57.9	59.8	40.0	50.7	57.6	62.7	44.2	55.5	31.4	39.9	16.5	32.4
ci_l	56.8	58.3	38.1	48.9	56.6	61.3	42.6	53.9	30.8	39.3	16.0	31.8
ci_h	59.0	61.3	41.6	52.5	58.7	64.0	45.8	57.2	32.1	40.6	17.1	32.9
Baseline + Captioning-to-VQA + Classification QA (*Q*Rad)												
median	58.3	59.5	41.5	51.8	57.9	62.2	45.1	55.2	31.6	40.2	16.7	32.5
ci_l	57.3	57.9	39.8	49.7	56.9	60.8	43.7	53.6	30.9	39.4	16.2	32.0
ci_h	59.4	61.0	42.97	53.7	59.0	63.5	46.6	57.0	32.2	40.9	17.2	33.1

[1.] The baseline ablates Captioning-to-VQA reframing, while keeping model architecture and pre-training the same. It is equivalent to the previous work in MedImageInsight (MI2) [11].
[2.] To demonstrate statistical significance, we report the median and 95% confidence intervals (ci_l and ci_h) over 500 bootstrap replicates for all metrics.

4 Conclusion

In this paper, we introduce *Q*Rad, a novel approach that reframes long text generation, such as radiology reports, from captioning to a chain of VQA process. Our problem reformulation improves the factual quality, enables user interaction, and allows probability-based evaluation such as ROC curves. *Q*Rad improves the clinical utility of report generation with 13% of the model size. Beyond radiology, our approach offers a novel framework for high-stakes domains where both factual correctness and interactive capabilities are essential.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, et al. MAIRA-2: Grounded Radiology Report Generation. *arXiv preprint arXiv:2406.04449*, 2024.

[4] Brent Burbridge. *Undergraduate diagnostic imaging fundamentals*. Distance Education Unit, University of Saskatchewan, 2017.

[5] Daniel J Cao, Casey Hurrell, and Michael N Patlas. Current status of burnout in canadian radiology. *Canadian Association of Radiologists Journal*, 74(1):37–43, 2023.

[6] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

[7] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Hassan Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu-Hsin Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. 2024.

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In *ECCV*, 2020.

[9] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[10] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

[11] Noel C. F. Codella, Ying Jin, Shrey Jain, Yu Gu, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Naiteek Sangani, Sheng Zhang, Hoifung Poon, Stephanie Hyland, Shruthi Bannur, Javier Alvarez-Valle, Xue Li, John Garrett, Alan McMillan, Gaurav Rajguru, Madhu Maddi, Nilesh Vijayrania, Rehaan Bhimai, Nick Mecklenburg, Rupal Jain, Daniel Holstein, Naveen Gaur, Vijay Aski, Jenq-Neng Hwang, Thomas Lin, Ivan Tarapov, Matthew Lungren, and Mu Wei. Medimageinsight: An open-source embedding model for general domain medical imaging, 2024.

[12] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.

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

[14] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. *arXiv preprint arXiv: 2111.02387*, 2021.

[15] US Food. Drug administration. statistical guidance on reporting results from studies evaluating diagnostic tests-guidance for industry and fda staff. *Food and Drug Administration, Center for Devices and Radiological Health Diagnostic Devices Branch, Division of Biostatistics, Office of Surveillance and Biometrics*, 2007.

[16] Dhakshinamoorthy Ganeshan, Andrew B Rosenkrantz, Roland L Bassett Jr, Lori Williams, Leon Lenchik, and Wei Yang. Burnout in academic radiologists in the united states. *Academic radiology*, 27(9):1274–1281, 2020.

[17] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Stack-captioning: Coarse-to-fine learning for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

[18] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 503–519, 2018.

[19] Xinyue Hu, L Gu, Q An, M Zhang, L Liu, K Kobayashi, T Harada, R Summers, and Y Zhu. Medical-diff-vqa: a large-scale medical dataset for difference visual question answering on chest x-ray images. *PhysioNet*, 12:13, 2023.

[20] Xinyue Hu, Lin Gu, Kazuma Kobayashi, Liangchen Liu, Mengliang Zhang, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. Interpretable medical image visual question answering via multi-modal relationship graph learning. *Medical Image Analysis*, 97:103279, 2024.

[21] Chen Huang, Shuangfei Zhai, Walter Talbott, Miguel Bautista Martin, Shih-Yu Sun, Carlos Guestrin, and Josh Susskind. Addressing the loss-metric mismatch with adaptive loss alignment. In *International conference on machine learning*, pages 2891–2900. PMLR, 2019.

[22] Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. MAIRA-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.

[23] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019.

[24] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

[25] Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 105–115, 2021.

[26] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2607–2615, 2024.

[27] Haibo Jin, Haoxuan Che, Sunan He, and Hao Chen. A chain of diagnosis framework for accurate and explainable radiology report generation. *IEEE Transactions on Medical Imaging*, 2025.

[28] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

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

[30] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[31] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[32] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.

[33] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLava-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[36] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[38] Hector Lopez-Cardona. Chest x-ray review: Abcde. *Radiopaedia*, 2023.

[39] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*, 2020.

[40] Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yucheng Tang, Pengfei Guo, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. *arXiv preprint arXiv:2411.12915*, 2024.

[41] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *European Conference on Computer Vision*, pages 167–184. Springer, 2022.

[42] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest X-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633, 2023.

[43] Aaron Nicolson, Jinghui Liu, Jason Dowling, Anthony Nguyen, and Bevan Koopman. e-health CSIRO at RRG24: Entropy-augmented self-critical sequence training for radiology report generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 99–104, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[44] Aaron Nicolson, Shengyao Zhuang, Jason Dowling, and Bevan Koopman. The impact of auxiliary patient data on automated chest X-ray report generation and how to incorporate it. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–203, Vienna, Austria, 2025. Association for Computational Linguistics.

[45] Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. *arXiv preprint arXiv:2102.09777*, 2021.

[46] OpenAI. Gpt-4 technical report, 2023.

[47] Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikannan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. Rexvqa: A large-scale visual question answering benchmark for generalist chest x-ray understanding. *arXiv preprint arXiv:2506.04353*, 2025.

[48] Jay R Parikh, Darcy Wolfman, Claire E Bender, and Elizabeth Arleo. Radiologist burnout according to surveyed radiology practice leaders. *Journal of the American College of Radiology*, 17(1):78–81, 2020.

[49] Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Benedikt Wiestler, Nassir Navab, and Matthias Keicher. Radialog: Large vision-language models for x-ray reporting and dialog-driven assistance. In *Medical Imaging with Deep Learning*, 2025.

[50] Mercy Ranjit, Shaury Srivastav, and Tanuja Ganu. Radphi-3: Small language models for radiology. *arXiv preprint arXiv:2411.13604*, 2024.

[51] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[52] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

[53] Robin Smithuis and Delden Otto. Chest x-ray - basic interpretation. *Radiology Assistant*, 2022.

[54] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023.

[55] Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, et al. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608, 2025.

[56] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards Generalist Biomedical AI. *NEJM AI*, 1(3): AIoa2300138, 2024.

[57] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[58] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. UFO: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021.

[59] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.

[60] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.

[61] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[62] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[63] Benjamin Yan, Ruochen Liu, David E Kuo, Subathra Adithan, Eduardo Pontes Reis, Stephen Kwak, Vasantha Kumar Venugopal, Chloe P O'Connell, Agustina Saenz, Pranav Rajpurkar, et al. Style-aware radiology report generation with radgraph and few-shot prompting. *arXiv preprint arXiv:2310.17811*, 2023.

[64] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024.

[65] Yanzhi Yi, Hangyu Deng, and Jinglu Hu. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994, 2020.

[66] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[67] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[68] Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. Libra: Leveraging temporal images for biomedical radiology analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17275–17303, Vienna, Austria, 2025. Association for Computational Linguistics.

[69] Hong-Yu Zhou, Julián Nicolás Acosta, Subathra Adithan, Suvrankar Datta, Eric J Topol, and Pranav Rajpurkar. Medversa: A generalist foundation model for medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.

A Implementation Details and Evaluation Metrics

Training Parameters. We use an image size of 512x512. During the pre-training on CXR-697K (consistent with existing work LLaVA-Rad [7]), we freeze the image encoder and updates the text decoder with a learning rate of 2E-5 for 400 epochs. The batch size is 2048, and no instruction is used in this phase. We then fine-tune the model on MIMIC-CXR using the official training split with a learning rate of 1E-5 for the image encoder and 5E-5 for the text decoder. The training takes 5 hours with 128 V100 GPUs, using a batch size is 512 and set for 60 epochs. We mixed the report generation QA and image classification QA by a ratio of 6:4.

Evaluation Metrics. CheXbert [52] is a Clinical Efficacy (CE) metric that classifies generated reports into 14 disease categories and evaluates them based on classification performance, focusing on factual accuracy rather than textual overlap. RadGraph [24] is designed specifically for radiology reports and assesses the correctness of extracted entities and their attributes. BLEU and ROUGE are standard lexical metrics that measure n-gram similarity to evaluate text overlap. We use results from [7] if not available in the original papers.

B Qualitative Results

Qualitative examples illustrating the *Q*Rad pipeline are shown in Figure 4. The figure highlights three key aspects:

- **Intrinsic VQA capability:** The predicted answers are directly relevant to the input questions, demonstrating the model's ability to perform visual question answering.
- **Factual correctness:** The model generates factually accurate answers, although there may be stylistic differences such as sentence structure or order.
- **Interactive refinement:** When provided with ground-truth questions (simulating a scenario where a radiologist requests specific information), the model produces answers that are both reasonable and closely aligned with the ground-truth responses.

C ROC-based Evaluation for Regulatory Validation

*Q*Rad is the first report generation model to produce class probabilities scores for defined disease directly from its text generation components. Unlike multi-task models that use separate modules for classification and text generation, *Q*Rad generates both outputs from the same component. This design enables the evaluation of confidence scores to directly reflect the model's intrinsic classification capability.

The class probabilities are used to generate Receiver Operating Characteristic (ROC) and sensitivity-specificity curves, which are typical in FDA approval studies for diagnostic systems. As shown in Figure 5, this confidence-based evaluation provides more granular insights into clinical utility, such as the characteristics of the sensitivity and specificity trade-off. This is especially valuable because clinical applications often have different costs for false positives and false negatives.

From Figure 5, we observe that *Q*Rad performs reliably on classes such as Enlarged Cardiomediastinum, Cardiomegaly, and Lung Opacity, but is less reliable on Pleural Other, Fracture, and Pneumothorax. We attribute this difference to two main factors. First, conditions like fracture require detection of subtle details and are rare in the dataset. Second, some classes (e.g., Pleural Other) aggregate many rare disease names, making it challenging for our prompts to comprehensively elicit the expected output.

The ROC curve enables more comprehensive guidance for clinical adoption by illustrating the model's characteristics across different sensitivity-specificity operating points, rather than relying solely on binary predictions as in existing CheXbert-based metrics. For example, in a copilot system that alerts radiologists to potential missed findings, maximizing sensitivity may be prioritized to ensure that as few true cases as possible are overlooked. Conversely, in automated triage systems that escalate only the most critical or certain cases for urgent review, higher specificity may be preferred to avoid unnecessary interruptions and reduce alarm fatigue. The ROC curve allows stakeholders to evaluate

(a) *Q*Rad Predicted Questions-Answer Pairs

- [Q1] "What is the position of the endotracheal tube?"
  [A1] "The endotracheal tube is 3.5 cm above the carina."
- [Q2] "Where does the right-sided picc line terminate?"
  [A2] "A right-sided picc line terminates in the mid svc."
- [Q3] "What is the course of the enteric tube?"
  [A3] "An enteric tube courses below the diaphragm with the tip out of the field of view."
- [Q4] "Are there any changes in the lung opacities compared to previous studies?"
  [A4] "Since the prior study, there has been improvement in the bilateral parenchymal opacities. "
- [Q5] "What is the status of the lung volumes and the cardiomediastinal silhouette?"
  [A5] "Lung volumes are low and the cardiomediastinal silhouette is enlarged, but stable."

(b) *Q*Rad Prediction Compared with Ground-truth

- [GT-Q1] "Are there any support devices present?"
  [GT-A1] "An endotracheal tube, NG tube, and right upper extremity PICC with its tip at the cavoatrial junction are unchanged."
  [Pred-A1] "An endotracheal tube is in place with the tip terminating 3.5 cm above the carina. A nasogastric tube is seen coursing below the diaphragm with the tip terminating in the left upper quadrant. A right picc is unchanged in position with the tip terminating in the mid-to-low svc."
- [GT-Q2] "Is there any change in lung opacity?"
  [GT-A2] "There is no change in left lower lobe opacity."
  [Pred-A2] "There has been interval improvement in the right basilar opacity."
- [GT-Q3] "Are there signs of pleural effusion or pneumothorax?"
  [GT-A3] "There is no large pleural effusion, or pneumothorax."
  [Pred-A3] "No pleural effusion or pneumothorax is detected. "
- [GT-Q4] "What is the condition of the cardiac silhouette and mediastinal contours?"
  [GT-A4] "The cardiac silhouette remains moderately enlarged, mediastinal contours are notable for calcification of the aortic arch."
  [Pred-A4] "The cardiac silhouette is mildly enlarged but stable, the mediastinal contours are prominent but stable."

Figure 4: Qualitative example of question-answer generation by *Q*Rad. (a) displays a set of questions (Q_n) generated by *Q*Rad for a given image, along with their corresponding answers (A_n). (b) compares GPT-generated questions (GT-Q_n), ground-truth answers from the reference report (GT-A_n), and *Q*Rad's predicted answers (Pred-A_n) for each question. *Q*Rad demonstrates factually reliable outputs, even if the order of information differs.

the model's behavior on disease classes relevant to the clinical context and risk tolerance, thereby assessing its practical utility more faithfully.

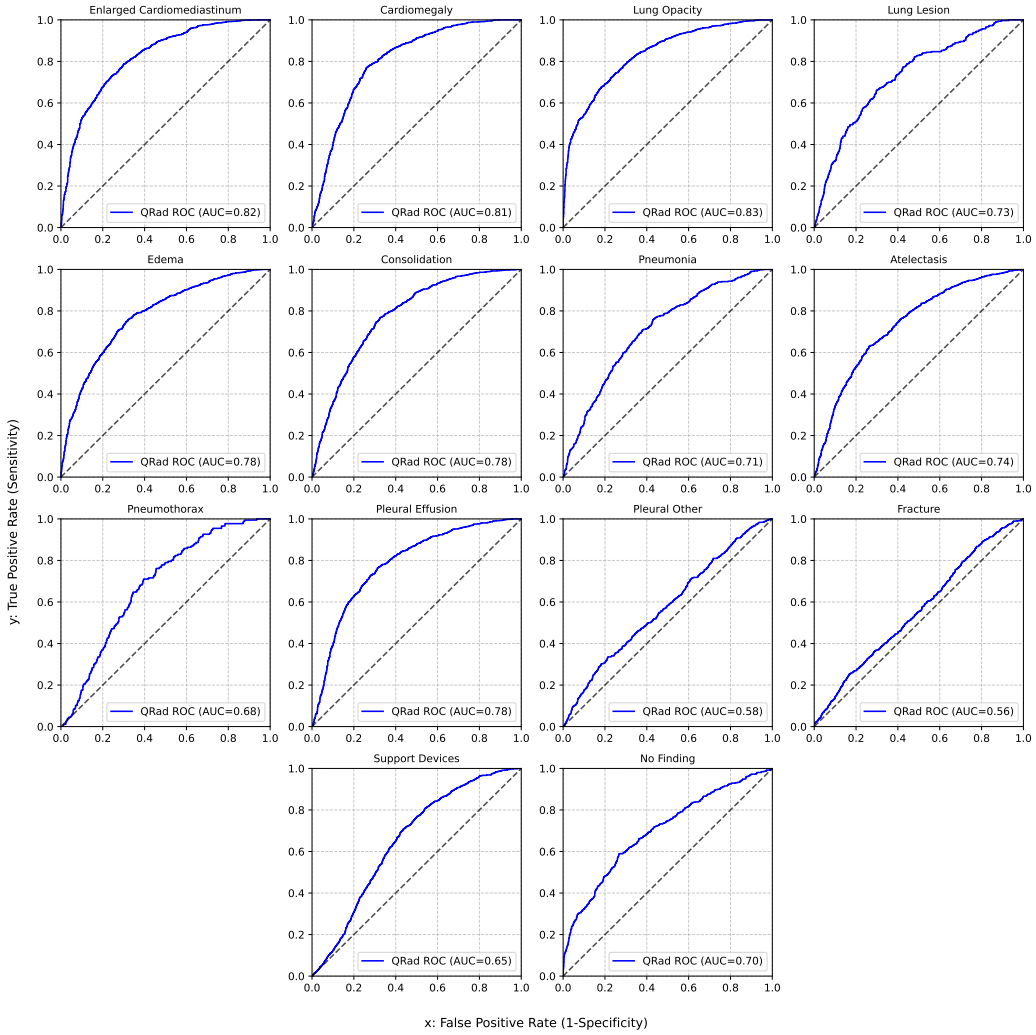

Figure 5: Receiver Operating Characteristic (ROC) curves for *Q*Rad across multiple disease classes. The x-axis shows the False Positive Rate (1-specificity), and the y-axis shows the True Positive Rate (sensitivity). Each curve illustrates the trade-off between sensitivity and specificity. The ROC analysis enables a nuanced assessment of *Q*Rad's clinical utility across different operating points and disease categories.

D Results on IU X-ray

Following recent studies [7, 3], we use the IU X-ray dataset [13] as a fully held-out evaluation set. All 3198 frontal-view X-rays are used as the testing split unseen during training. Results in Appendix D shows that *Q*Rad generates well on unseen data.

E Additional Tables for Ablation Study and Hyper-parameters

We provide additional tables for ablation studies and hyper-parameters such as dataset mixture ratios in Table 5 and Table 6.

Table 4: Report Generation Performance on IU-XRay

Model	CheXbert ("uncertain" as *negative*)				CheXbert ("uncertain" as *positive*)				RadGraph	BLEU		ROUGE
	Micro-avg		Macro-avg		Micro-avg		Macro-avg					
	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	ER	(1)	(4)	(L)
R2Gen [9]	-	-	13.6	-	-	-	-	-	-	32.5	5.9	25.3
CvT2Dist. [42]	-	-	16.8	-	-	-	-	-	-	38.3	8.2	**27.7**
RGRG [54]	-	-	18.0	-	-	-	-	-	-	26.6	6.3	18.0
LLaVA-Rad [7]	**53.5**	-	-	-	-	-	-	-	-	-	-	25.3
*Q*Rad	46.5	36.9	**27.0**	27.2	44.3	38.8	28.7	31.2	29.4	**41.9**	**10.8**	25.3

F Ethical Considerations

Medical datasets often contain sensitive patient information. To ensure the ethical use of such data, this study adheres to strict guidelines. All participants who accessed the MIMIC-CXR dataset, including the authors and radiologists involved in this research, completed the required onboarding process through PhysioNet[3]. For the IU X-ray dataset, we complied with the license[4].

To maintain compliance with PhysioNet's policy on the use of large language model APIs during the automatic evaluation, we utilized a secure, private, in-house deployment of GPT-4. This approach guarantees that no sensitive information is shared with external parties.

Furthermore, to protect patient privacy, X-ray images presented in this paper were carefully selected from open, compliance-free sources, ensuring that no identifiable patient information is disclosed.

[3]MIMIC-CXR on PhysioNet: https://physionet.org/content/mimic-cxr/2.0.0/

[4]IU X-ray dataset license: https://creativecommons.org/licenses/by-nc-nd/4.0/

Table 5: Comparison of Dataset Hyper-parameters on MIMIC-CXR

Model	CheXbert ("uncertain" as *negative*)				CheXbert ("uncertain" as *positive*)				RadGraph	BLEU		ROUGE
	Micro-avg		Macro-avg		Micro-avg		Macro-avg					
	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	ER	(1)	(4)	(L)
(a) Classification QA = 20%, Label={P}												
median	57.9	59.6	40.4	51.2	57.4	61.8	44.6	54.7	31.4	40.5	16.6	32.4
ci_l	56.8	57.8	38.8	49.1	56.4	60.5	43.1	53.2	30.7	39.8	16.1	31.8
ci_h	59.0	61.1	42.0	53.1	58.5	63.3	46.2	56.4	32.0	41.2	17.2	32.9
(b) Classification QA = 40%, Label={P}												
median	57.8	59.5	40.2	50.9	57.3	61.7	44.5	54.4	31.4	40.0	16.6	32.5
ci_l	56.7	58.0	38.6	48.9	56.4	60.2	43.0	52.9	30.8	39.3	16.1	31.9
ci_h	58.9	61.0	41.9	52.7	58.4	63.1	46.2	56.2	32.1	40.8	17.1	33.1
(c) Classification QA = 40%, Label={P, Random U}												
median	57.9	59.2	40.5	50.8	57.5	61.6	44.5	54.5	31.3	40.1	16.6	32.4
ci_l	57.0	57.7	38.8	48.8	56.5	60.3	42.9	53.0	30.7	39.4	16.1	31.9
ci_h	59.1	60.7	42.0	52.6	58.5	63.1	46.0	56.3	31.9	40.8	17.2	33.0
(d) Classification QA = 40%, Label={P, U}, w/o QA Context												
median	56.6	58.8	39.6	50.7	56.3	61.3	43.4	54.1	28.0	41.8	13.8	27.9
ci_l	55.5	57.4	38.2	48.7	55.4	59.9	42.1	52.6	27.5	41.3	13.5	27.5
ci_h	58.0	60.3	41.1	52.5	57.4	62.7	44.8	55.8	28.6	42.3	14.2	28.3
(e) Classification QA = 40%, Label={P, U}												
median	58.3	59.5	41.5	51.8	57.9	62.2	45.1	55.2	31.6	40.2	16.7	32.5
ci_l	57.3	57.9	39.8	49.7	56.9	60.8	43.7	53.6	30.9	39.4	16.2	32.0
ci_h	59.4	61.0	42.97	53.7	59.0	63.5	46.6	57.0	32.2	40.9	17.2	33.1

[1.] The Classification QA ratio (0%, 20%, 40%) indicates the proportion of classification QA pairs in the training data.
[2.] The Label field defines which CheXbert classes are mapped to the positive classes in the Classification QA data: P (positive only), U (positive and uncertain), Random U (uncertain randomly used as positive)
[3.] QA Context represents that whether previous QA pairs are provided as input context. The w/o QA Context ablates this feature.

Table 6: Ablation of the Question Generator

Model	CheXbert ("uncertain" as *negative*)				CheXbert ("uncertain" as *positive*)				RadGraph	BLEU		ROUGE
	Micro-avg		Macro-avg		Micro-avg		Macro-avg					
	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	ER	(1)	(4)	(L)
Template Questions [a]	44.7	52.1	28.5	37.5	46.6	56.3	34.2	45.5	23.5	31.1	8.2	20.9
Predicted Questions [b]	**58.4**	**59.5**	**41.5**	**51.8**	**57.9**	**62.2**	**45.1**	**55.2**	**31.5**	**40.0**	**16.9**	**32.5**
Oracle Questions [c]	74.7	78.0	60.5	72.2	76.1	79.8	66.0	74.9	48.0	54.4	30.5	52.8

[a.] We use the template questions for all input images composed from the 14 widely used CheXbert classes
[b.] We use the Question Generator to learn and predict the questions per input image (the *Q*Rad method)
[c.] Oracle questions are ChatGPT-generated directly from the ground truth reports. This is the ground truth used to train the Question Generator