CURV: Coherent Uncertainty-Aware Reasoning in Vision-Language Models for X-Ray Report Generation

Ziao Wang[1,2], **Sixing Yan**[1], **Kejing Yin**[1,*], **Xiaofeng Zhang**[3], **William K. Cheung**[1]

[1]Department of Computer Science, Hong Kong Baptist University
[2]Institute of Systems Medicine and Health Sciences, Hong Kong Baptist University
[3]Department of Computer Science, Harbin Institute of Technology

Abstract

Vision-language models have been explored for radiology report generation with promising results. Yet, uncertainty elaborated in findings and the reasoning process for reaching clinical impressions are seldom explicitly modeled, reducing the clinical accuracy and trustworthiness of the generated reports. We present CURV, a novel framework that alleviates the limitations through integrated awareness of uncertainty and explicit reasoning capabilities. Our approach consists of three key components: (1) an uncertainty modeling mechanism that teaches the model to recognize and express appropriate levels of diagnostic confidence, (2) a structured reasoning framework that generates intermediate explanatory steps connecting visual findings to clinical impressions, and (3) a reasoning coherence reward that ensures logical consistency among findings, reasoning, and impressions. We implement CURV through a three-stage training pipeline that combines uncertainty-aware fine-tuning, reasoning initialization, and reinforcement learning. In particular, we adopt a comprehensive reward function that addresses multiple aspects of report quality, incorporating medical term matching, uncertainty expression evaluation, and semantic coherence evaluation. Experimental results demonstrate that CURV generates clinically relevant reports with appropriate uncertainty expressions and transparent reasoning traces, significantly outperforming previous methods. CURV[2] represents a substantial advancement toward interpretable and trustworthy AI-generated radiology reports, with broader implications for the deployment of vision-language models in high-stakes clinical environments where uncertainty awareness and reasoning transparency are essential.

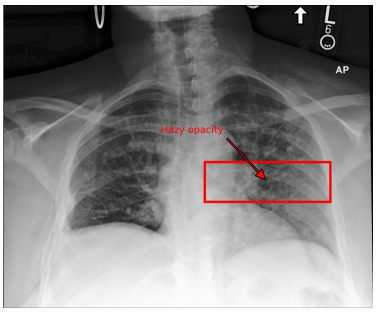
1 Introduction

Chest X-rays (CXRs) are a cornerstone of diagnostic imaging, yet their interpretation is time-intensive, straining healthcare systems amid radiologist shortages. Automated report generation using vision-language models (VLMs) offers a promising solution to enhance efficiency and reduce workload [31, 20]. Moreover, this technology can help bridge the gap in report quality between large and small hospitals, where physicians in smaller facilities often have less experience compared to those in major centers, enabling more consistent and accurate diagnoses across diverse settings.

Medical report generation, unlike general image captioning tasks, imposes unique challenges that current VLMs are not fully equipped to address, particularly in handling diagnostic uncertainty and providing transparent reasoning [35]. Radiologists routinely employ linguistic markers such as "likely"

[*]Correspondence to: Kejing Yin <cskjyin@comp.hkbu.edu.hk>
[2]Code and Dataset: `https://github.com/wwwadx/CURV`

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Chest X-ray image

Structural Uncertainty (Findings): *"Pulmonary nodules in the left upper lobe are also **not completely characterized** on this study. However, in addition, there is a more hazy widespread opacity projecting over the left mid upper lung which **could be compatible with** a coinciding pneumonia."*

Semantic Uncertainty (Impression): *"Increasing left lung opacification which **may reflect** pneumonia superimposed on metastatic disease, although other etiologies such as lymphangitic pattern of metastatic spread **could be considered**. CT **may be helpful** to evaluate further if needed clinically."*

(b) Corresponding uncertain expressions from the radiology report.

Figure 1: Illustration of Structural and Semantic Uncertainty. (a) A hazy opacity in the left mid-upper lung (arrow) generates structural uncertainty regarding its nature. (b) This feeds into the semantic uncertainty in the impression, where multiple etiologies are considered and further investigation (CT) is suggested.

or "possible" to convey varying degrees of diagnostic confidence, ensuring clear communication with referring physicians [32]. As diagnostic uncertainty is multifaceted, as illustrated in Figure 1, it is crucial to distinguish between at least two key types: (1) **Structural Uncertainty** regarding specific visual findings (e.g., a sentence in the "Findings" section stating "hazy opacity... *could be compatible with* pneumonia"); and (2) **Semantic Uncertainty** stated in the overall "Impression" (e.g., "opacification *may reflect* pneumonia... although other etiologies *could be* considered"). Accurately modeling these uncertainties, and the reasoning that connects them, is crucial for generating trustworthy reports to support effective clinical decision-making. Crucially, our work frames this challenge not as quantifying a model's internal statistical confidence, but as modeling the linguistic expression of diagnostic uncertainty—the specific language radiologists use to convey confidence levels. Moreover, robust clinical decision-making relies heavily on explicit reasoning that logically connects such visual findings, with their structural uncertainties, to the more diagnosis-oriented medical impressions and their associated semantic uncertainties.

Existing VLMs for CXR report generation often prioritize factual accuracy over modeling diagnostic uncertainty or providing explicit reasoning pathways, resulting in reports that lack clinical nuance and transparency [41, 29, 8, 38, 23]. This deficiency poses a significant barrier to their clinical adoption, as physicians require both accurate uncertainty expression and transparent reasoning to trust and effectively utilize AI-generated reports. Consequently, there is a pressing need for developing approaches that address these critical gaps by integrating robust uncertainty awareness and transparent reasoning mechanisms into the report generation process.

To tackle these challenges, we introduce CURV, a novel framework for uncertainty-aware vision-language models with explicit reasoning capabilities for CXR report generation. CURV advances the field through three key innovations: (1) a systematic uncertainty modeling approach that enables the model to recognize anatomical structures and express appropriate diagnostic confidence using a data-driven fine-tuning strategy and a specialized uncertainty reward; (2) a structured reasoning framework that generates intermediate explanatory steps linking visual findings to clinical impressions, thereby enhancing transparency. To enable the development and supervised initialization of this reasoning capability, we created **TRACE-CXR** (Transparent Reasoning and Articulation for Clinical Explanations - CXR), a novel dataset of 2,000 chest X-ray reports, each augmented with an explicit, LLM-generated "thinking" section that models the reasoning pathway from findings to impression; and (3) a multi-dimensional reward design in reinforcement learning that ensures logical consistency across findings, reasoning, and impressions, addressing the need for coherent and trustworthy reports. Our approach leverages a vision language model, augmented by a meticulously designed training pipeline that embeds uncertainty awareness and reasoning capabilities directly into the model. The uncertainty modeling component is developed through targeted fine-tuning with uncertainty-annotated data, enabling the model to predict confidence levels for identified pathologies and map them to appropriate linguistic expressions in the generated text. For the reasoning component, we employ a structured generation process that articulates visual findings, produces intermediate reasoning steps, and delivers clinical impressions—all while maintaining logical coherence across these elements through a novel optimization strategy.

The contributions of this study are as follows:

- We propose a framework for uncertainty-aware medical report generation, integrating a specialized fine-tuning strategy with curated uncertainty-annotated data and uncertainty-calibrated reward mechanism to enhance the clinical relevance of AI-generated CXR reports.
- We introduce a structured reasoning framework that leverages our **TRACE-CXR** dataset to initialize explicit "thinking" pathways within a tripartite report structure (findings, thinking, impression), supported by a multi-dimensional reward design in reinforcement learning, to ensure transparent explanations and logical consistency between radiological observations and clinical impressions.
- Through extensive experimentation, we demonstrate that CURV generates clinically relevant CXR reports with appropriate uncertainty expressions and transparent reasoning traces, outperforming existing methods in both qualitative and quantitative metrics, thus advancing trustworthy AI in high-stakes clinical environments.

2 Related Work

2.1 Uncertainty in Medical Report Generation

Expressing diagnostic uncertainty is critical in radiology for clear communication and effective clinical decision-making. Equipping vision-language models (VLMs) with uncertainty awareness is thus essential for clinical adoption. Existing studies have approached this from various perspectives. Wang et al. [32] used Monte Carlo dropout to estimate visual and textual uncertainty, integrating it into a weighted loss function for reliable outputs. Similarly, Yan et al. [35] introduced the Diagnostic Uncertainty Encoding framework to encode clinically inspired uncertainty concepts, enhancing report accuracy. Najdenkoska et al. [22] proposed a probabilistic latent variable model with variational topic inference to generate diverse CXR reports reflecting multiple interpretations. Additionally, Yang et al. [37] demonstrated that training Bayesian neural networks with uncertain labels increases predictive variance for ambiguous cases, while large-scale VLMs like Med-Gemini [25] employ uncertainty-guided strategies for clinical reasoning tasks. However, many approaches lack explicit differentiation between uncertainty in specific findings (Structural Uncertainty) and overall diagnostic synthesis (Semantic Uncertainty), as noted by [41, 29]. CURV addresses this gap by systematically modeling and expressing both types of uncertainty through a specialized reward mechanism and fine-tuning with uncertainty-annotated data, aiming for clinically nuanced reports.

2.2 Reasoning by Reinforcement Learning

Reinforcement learning (RL) has shown promise in enhancing reasoning capabilities in language and vision-language models [40, 18, 34]. DeepSeek-R1 [7] leverages Group Relative Policy Optimization (GRPO) [28] to develop reasoning skills via rule-based rewards. Extending RL to VLMs, Huang et al. [12] introduced Vision-R1, achieving strong performance through automated dataset construction and Progressive Thinking Suppression Training with GRPO [26, 39]. In the medical domain, RL is increasingly applied to radiology report generation for structured reasoning and trustworthiness [11, 42]. Jing et al. [15] proposed BoxMed-RL, focusing on explainable reports through CoT reasoning and spatially verifiable RL, emphasizing spatial grounding via IoU-based rewards. Similarly, Shao et al. [27] used RL to improve alignment between radiology images and reports, targeting linguistic quality and anomaly detection. While these works highlight RL's potential for logical consistency and transparency, CURV uniquely integrates uncertainty modeling with structured reasoning, using a multi-dimensional reward design to ensure coherent and trustworthy AI-generated reports.

3 The Proposed Method

The CURV framework, detailed in this section, is designed to produce radiology reports that embody both diagnostic accuracy and a sophisticated understanding of clinical uncertainty. Our method explicitly models the *Structural Uncertainty* (tied to individual findings) and *Semantic Uncertainty* (concerning the overall diagnostic impression), along with the coherent reasoning that bridges them.

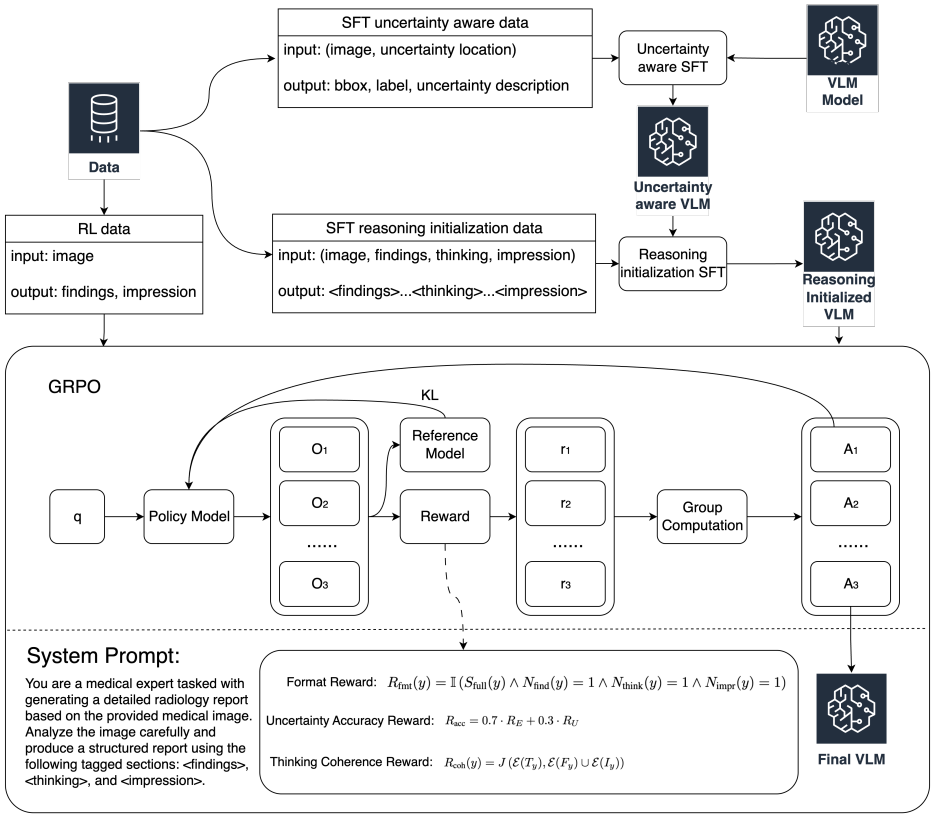


Figure 2: The CURV framework architecture. The figure illustrates the three-stage training pipeline, beginning with SFT for uncertainty awareness, followed by reasoning initialization, and culminating in a Reinforcement Learning phase using GRPO to refine the final Vision-Language Model.

3.1 Problem Formulation

The generation of medical reports from chest X-ray (CXR) images using vision-language models (VLMs) is a complex task requiring diagnostic accuracy, uncertainty expression, and transparent reasoning. Formally, given a CXR image I and a prompt p (e.g., requesting a structured report), the goal is to train a VLM π_θ to generate a structured output y with three components: visual findings, logical reasoning, and clinical impressions with uncertainty expressions. The objective is to optimize report quality across accuracy, uncertainty awareness, and coherence, formulated as:

$$\theta^* = \arg\max_\theta \mathbb{E}_{(I,p)\sim\mathcal{D}, y\sim\pi_\theta}[r(y, I, p)], \tag{1}$$

where $\mathcal{D}$ is the dataset of image-prompt pairs, y is the generated report, and $r(y, I, p)$ is a multi-dimensional reward function assessing format adherence, medical accuracy, uncertainty expression, and reasoning coherence. This framework, CURV, aims to produce clinically relevant, transparent, and trustworthy AI-generated reports by addressing these critical dimensions in a unified manner.

3.2 Empowering Uncertainty Awareness in Vision-Language Models

Current vision-language models (VLMs) often fail to capture the inherent uncertainty in medical report generation, a critical limitation for clinical decision-making in radiology. The CURV framework addresses this gap by instilling awareness of *Structural Uncertainty*, which pertains to specific findings, as introduced in Section 1. This initial stage focuses on enabling the model to recognize and articulate appropriate confidence levels for anatomical abnormalities, thereby enhancing the reliability of the "Findings" section in generated reports. The supervised fine-tuning (SFT) process leverages a curated uncertainty-annotated dataset $\mathcal{D}$, where each instance includes a CXR image I, a prompt p (e.g., instructing detection of anatomical objects with uncertainties), and ground-truth structural

uncertainty specifications $Y_{gt} = \langle f_1, f_2, \ldots, f_{N_I} \rangle$. Each finding $f_k = (b_k, l_k, t_k)$ encapsulates a bounding box, anatomical label, and textual uncertainty description. The model π_θ is trained to generate a serialized token sequence $\text{seq}(Y_{gt})$ using the loss function:

$$\mathcal{L}_{\text{uncertainty}} = - \sum_{(I,p,Y_{gt}) \in \mathcal{D}} \log \pi_\theta(\text{seq}(Y_{gt}) | I, p), \tag{2}$$

where $\log \pi_\theta(\text{seq}(Y_{gt}) | I, p)$ is the log-probability of producing the target sequence encoding all structural findings. This loss guides the model to detect uncertain regions (via b_k), identify them (via l_k), and articulate specific *Structural Uncertainty* (via t_k) in a structured, sequentially generated format.

3.3 Reasoning Initialization for Transparent Reporting

Building on the uncertainty awareness developed earlier, the second stage of CURV focuses on initializing basic reasoning capabilities in the model. The objective is to generate transparent reports by articulating logical connections between radiological findings and clinical impressions through an intermediate reasoning path, enhancing interpretability for clinical validation.

To achieve this, a report generation task is structured with three components: Findings (F), Reasoning (R), and Impression (C), with R elucidating the transition $F \xrightarrow{R} C$ to mimic a radiologist's inferential steps. Inspired by frameworks like DeepSeek-R1 [7], which use structured data to improve reasoning patterns, this stage employs supervised fine-tuning with a dataset we developed for this purpose, $\mathcal{D}_{\text{reason}}$. Each instance includes a CXR image I, a prompt p (e.g., "Generate a detailed radiology report"), and a ground-truth structured report $Y_{\text{structured}} = \langle \text{text}_F, \text{text}_R, \text{text}_C \rangle$. Here, text_R provides a logical narrative linking observations in text_F to conclusions in text_C, detailing abnormalities, suspected conditions, differential diagnoses, and contextual information. The VLM π_θ is fine-tuned on $\mathcal{D}_{\text{reason}}$ to maximize the log-likelihood of the serialized sequence $\text{seq}(Y_{\text{structured}})$, with the loss defined as:

$$\mathcal{L}_{\text{reasoning}} = - \sum_{(I,p,Y_{\text{structured}}) \in \mathcal{D}_{\text{reason}}} \log \pi_\theta(\text{seq}(Y_{\text{structured}}) | I, p). \tag{3}$$

By training the model with $\mathcal{L}_{\text{reasoning}}$, we instill a foundational capability to generate reports that are not only descriptive but also explanatory, paving the way for more sophisticated coherence and alignment in the subsequent reinforcement learning stage.

3.4 Enhancing Clinical Reasoning with Reinforcement Learning

Building upon the foundational capabilities empowered by the previous stages for uncertainty awareness and structured reasoning, the final CURV stage employs Reinforcement Learning (RL) to refine the model's clinical reasoning process. Our SFT-then-RL methodology is designed to move beyond simple imitation [6]. The preceding SFT stage uses the TRACE-CXR dataset to teach the model the basic tripartite report structure, providing a foundational policy for exploration. In this RL phase, however, the model is trained on the MIMIC-CXR dataset and is provided with only the ground-truth findings and impression sections. This design avoids the "imitation trap" where models simply reproduce potentially flawed or suboptimal reasoning paths from the SFT data. Instead, it forces the model to discover a functionally coherent reasoning process on its own, guided only by the reward signals that measure the logical connection between the human-authored findings and impressions. This ensures the model learns to generate genuinely coherent reasoning rather than engaging in "pseudo reasoning" by merely mimicking a template [4].

To this end, we use Group Relative Policy Optimization (GRPO) [28], guided by a novel, multi-component reward function tailored for clinical report generation. A key advantage of this RL phase, particularly through the coherence reward (R_{coh}), is its ability to guide the model in generating the intermediate "thinking" process that logically connects findings to impressions, without requiring extensive supervised data for this specific reasoning component.

Given an input (I, p) and ground truth y_{gt}, GRPO samples G reports $\{y_g\}$ from policy $\pi_{\theta_{\text{old}}}$ and updates π_θ by maximizing:

$$J(\theta) = \mathbb{E}\left[\frac{1}{G} \sum_{g=1}^{G} \left(\min(\rho_g A_g, \text{clip}(\rho_g, 1-\epsilon, 1+\epsilon) A_g) \right) - \beta D_{KL}(\pi_\theta || \pi_{\text{ref}}) \right], \tag{4}$$

where $J(\theta)$ represents the objective function, which aims to maximize the expected reward for the policy π_θ while maintaining stability in training. The function incorporates a clipped advantage term to limit large policy updates and a KL-divergence penalty term (D_{KL}) with coefficient β to prevent excessive deviation from a reference policy π_{ref}. The expectation $\mathbb{E}$ is taken over sampled reports, with G denoting the number of sampled reports $\{y_g\}$ from the old policy $\pi_{\theta_{\text{old}}}$. And ρ_g is the importance sampling ratio and $A_g = R_{\text{total}}(y_g, y_{\text{gt}}) - \text{mean}(\{R_{\text{total}}(y_k, y_{\text{gt}})\})$ is the advantage. The total reward $R_{\text{total}}(y, y_{\text{gt}})$ is a weighted sum of three components:

$$R_{\text{total}}(y, y_{\text{gt}}) = w_{\text{fmt}} R_{\text{fmt}}(y) + w_{\text{acc}} R_{\text{acc}}(y, y_{\text{gt}}) + w_{\text{coh}} R_{\text{coh}}(y). \tag{5}$$

The total reward $R_{\text{total}}(y, y_{\text{gt}})$ evaluates the quality of a generated report y compared to the ground truth y_{gt} by combining three distinct reward components with corresponding weights w_{fmt}, w_{acc}, and w_{coh}. These components assess format adherence (R_{fmt}), medical accuracy and uncertainty alignment (R_{acc}), and reasoning coherence (R_{coh}), respectively, ensuring a comprehensive evaluation of clinical report quality.

Format Adherence Reward (R_{fmt}): A binary reward $\mathbb{I}(\cdot)$ verifying strict adherence to the tripartite structure ($\langle$findings$\rangle$, $\langle$thinking$\rangle$, $\langle$impression$\rangle$), ensuring each tag appears exactly once and in order, with no extraneous content.

$$R_{\text{fmt}}(y) = \mathbb{I}\left(S_{\text{full}}(y) \wedge N_{\text{find}}(y) = 1 \wedge N_{\text{think}}(y) = 1 \wedge N_{\text{impr}}(y) = 1\right), \tag{6}$$

where $R_{\text{fmt}}(y)$ is a binary indicator function $\mathbb{I}(\cdot)$ that returns 1 only if the generated report y fully adheres to the required structure. Specifically, $S_{\text{full}}(y)$ checks if all required sections are present, while $N_{\text{find}}(y) = 1$, $N_{\text{think}}(y) = 1$, and $N_{\text{impr}}(y) = 1$ ensure that each section (findings, thinking, impression) appears exactly once in the correct order.

Findings/Impression Uncertainty Accuracy Reward (R_{acc}): Evaluates medical accuracy and uncertainty in findings (F_y) and impression (I_y) sections against y_{gt}, weighted as $R_{\text{acc}} = 0.7 \cdot R_E + 0.3 \cdot R_U$. Returns 0 if sections are missing.

1. **Entity Matching (R_E):** Average F1-score of RadGraph-extracted entities [14] between F_y, I_y and $F_{y_{\text{gt}}}, I_{y_{\text{gt}}}$.

$$R_E(y, y_{\text{gt}}) = \frac{1}{2}\left(\text{F1}(\mathcal{E}(F_y), \mathcal{E}(F_{y_{\text{gt}}})) + \text{F1}(\mathcal{E}(I_y), \mathcal{E}(I_{y_{\text{gt}}}))\right), \tag{7}$$

   where $R_E(y, y_{\text{gt}})$ computes the average F1-score for entity matching between the generated report sections (F_y for findings and I_y for impression) and the ground truth sections ($F_{y_{\text{gt}}}$ and $I_{y_{\text{gt}}}$). The function $\mathcal{E}(\cdot)$ extracts entities using RadGraph, and $\text{F1}(\cdot, \cdot)$ measures the overlap between entity sets, equally weighting the performance on findings and impression sections.

2. **Uncertainty Alignment (R_U):** Average alignment of (entity, term, score) triples $\mathcal{P}(S_{y'})$ between sections.

$$R_U(y, y_{\text{gt}}) = \frac{1}{2}\left(\text{MatchPairs}(\mathcal{P}(F_y), \mathcal{P}(F_{y_{\text{gt}}})) + \text{MatchPairs}(\mathcal{P}(I_y), \mathcal{P}(I_{y_{\text{gt}}}))\right), \tag{8}$$

   where $R_U(y, y_{\text{gt}})$ measures the alignment of uncertainty expressions between the generated report sections (F_y and I_y) and the ground truth ($F_{y_{\text{gt}}}$ and $I_{y_{\text{gt}}}$). The function $\mathcal{P}(\cdot)$ extracts (entity, term, score) triples, and $\text{MatchPairs}(\cdot, \cdot)$ evaluates their similarity, averaging the results for findings and impression sections. MatchPairs combines semantic term similarity (0.4 weight) and score difference (0.6 weight) for common entities, then combines this average similarity (0.7 weight) with entity coverage (0.3 weight).

Thinking Coherence Reward (R_{coh}): Measures Jaccard similarity $J(\cdot, \cdot)$ between RadGraph-extracted entities in the thinking section $\mathcal{E}(T_y)$ and the union of entities in findings $\mathcal{E}(F_y)$ and impression $\mathcal{E}(I_y)$. Returns 0 if sections are missing or entity sets are empty.

$$R_{\text{coh}}(y) = J\left(\mathcal{E}(T_y), \mathcal{E}(F_y) \cup \mathcal{E}(I_y)\right), \tag{9}$$

where $R_{\text{coh}}(y)$ quantifies the coherence of the thinking section (T_y) in the generated report y by computing the Jaccard similarity $J(\cdot, \cdot)$ between entities extracted from T_y (via $\mathcal{E}(T_y)$) and the union of entities from the findings ($\mathcal{E}(F_y)$) and impression ($\mathcal{E}(I_y)$) sections. This ensures that the reasoning process logically connects observations to conclusions. This multifaceted RL optimization aims for reports that are structurally sound, factually accurate with appropriate uncertainty, and demonstrate transparent clinical reasoning.

4 Experiments and Analysis

4.1 Experimental Setup

We conducted experiments on 4xA100 GPUs using Qwen-2.5-VL-3B as the backbone model, balancing efficiency and performance. Training spanned multiple stages over approximately 100 hours with a batch size of 16 and a learning rate of 1×10^{-6}. CURV was benchmarked against established vision-language models like LLaVA-1.5-7B and MAIRA-2 under consistent conditions. Full details on configurations and baseline setups are provided in Appendix A.

4.2 Datasets

Our experiments leverage the MIMIC-CXR dataset. As a key contribution of this work, we curated specialized data subsets: (1) an uncertainty-annotated dataset with 112,111 samples, and (2) our novel **TRACE-CXR** dataset (**T**ransparent **R**easoning and **A**rticulation for **C**linical **E**xplanations in CXR), featuring 2,000 reports with explicit reasoning pathways. Both were developed to support uncertainty modeling and structured reasoning, respectively, significantly improving data utility. The clinical validity of our TRACE-CXR dataset was subsequently confirmed through a formal evaluation with a board-certified radiologist, which revealed a strong concordance with expert judgment (see Appendix B). Detailed curation processes and statistics for both datasets are also described in Appendix B.

4.3 Evaluation Metric

We assess CURV using standard NLP metrics (e.g., BLEU, ROUGE-L, METEOR) for textual quality and clinical accuracy metrics (e.g., CheXbert, RadGraph F1-scores) for medical relevance. Additionally, LLM-based evaluation protocols are employed to evaluate the unique "Thinking" section and uncertainty expressions (structural and semantic). Complete metric definitions and evaluation protocols are available in Appendix C.

Table 1: Generation metrics for radiology report generation across different models

Model	B-1	B-2	B-3	B-4	METEOR	R-L	gritlm
LLaVA-1.5-7B [21]	19.09	7.46	2.81	1.25	19.16	18.36	44.25
LLaVA-1.5-7B-SFT-CXR	22.58	15.06	9.43	6.13	25.71	28.09	50.28
HuatuoGPT-Vision-7B [5]	19.33	9.42	4.64	1.93	26.01	20.78	47.32
MAIRA-2 [3]	24.94	14.12	9.01	6.14	26.78	28.65	47.48
Qwen2.5-VL-3B [1]	13.09	5.42	2.08	0.89	20.81	15.23	44.57
Gemini 2.5 pro	12.54	5.20	2.25	1.05	21.19	15.01	40.41
CURV_stage1	14.96	7.08	3.33	1.61	23.47	19.07	45.15
CURV_stage2	10.72	5.22	2.43	1.10	18.57	14.59	42.76
CURV	**25.38**	**15.58**	**9.85**	**6.18**	**30.43**	**31.19**	**50.48**

Table 2: Clinical accuracy metrics for radiology report generation across different models

Model	CheXbert			RadGraph	
	Acc.	Macro F1	Micro F1	Ent. F1	F1
LLaVA-1.5-7B [21]	63.25	4.94	38.54	7.61	4.95
LLaVA-1.5-7B-SFT-CXR	72.72	5.00	51.51	17.57	13.06
HuatuoGPT-Vision-7B [5]	71.15	5.34	48.62	15.92	9.06
MAIRA-2 [3]	67.39	**6.34**	46.53	25.01	17.05
Qwen2.5-VL-3B [1]	67.78	4.75	37.66	9.46	4.66
Gemini 2.5 pro	74.35	5.35	48.45	13.09	7.71
CURV_stage1	57.59	4.51	30.75	17.00	9.95
CURV_stage2	56.57	3.87	26.83	11.11	6.03
CURV	**76.93**	5.22	**57.12**	**25.95**	**19.54**

4.4 Main Results

Overall Metric The experimental results in Tables 1 and 2 highlight the effectiveness of the CURV framework in radiology report generation, with notable insights from both metric interpretations and model scale perspectives. From the metrics' standpoint, CURV achieves superior generation quality with top scores in BLEU (e.g., BLEU-3: 9.85), METEOR (30.43), and ROUGE_L (31.19) compared to SOTA method like MAIRA-2 (BLEU-3: 9.01), indicating enhanced fluency and textual relevance. Likewise, CURV's ability to produce medically relevant content is underscored by its strong clinical accuracy. This advantage is particularly pronounced when compared to frontier generalist models; while such models are powerful, CURV's specialized approach significantly outperforms Gemini 2.5 Pro across all clinical metrics, achieving a RadGraph F1 score of 19.54 versus 7.71. This demonstrates the critical value of a targeted framework for this complex medical task. Finally, from a model scale perspective, CURV, with a compact 3B parameter size, outperforms larger 7B models like HuatuoGPT-Vision-7B and MAIRA-2, demonstrating that our proposed method enables efficient performance without requiring extensive computational resources.

Table 3: Generation metrics for radiology report generation on IU X-ray dataset

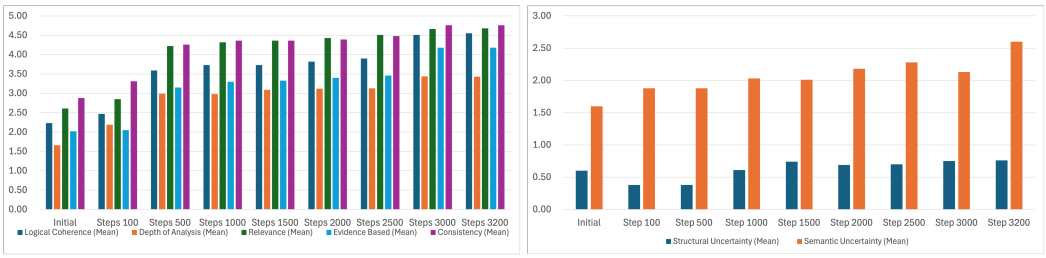
Model	**B-1**	**B-2**	**B-3**	**B-4**	**METEOR**	**R-L**	**gritlm**
LLaVA-1.5-7B	16.52	6.60	3.00	1.40	19.65	17.48	45.78
LLaVA-1.5-7B-SFT-CXR	21.42	12.95	8.03	5.20	23.24	26.40	46.21
HuatuoGPT-Vision-7B	19.33	10.70	6.28	2.81	31.02	23.42	50.22
MAIRA-2	26.37	15.60	9.64	6.03	25.52	31.18	54.18
Qwen2.5-VL-3B	11.38	5.05	2.28	1.05	21.65	15.01	45.95
CURV_stage1	12.51	6.24	3.18	1.57	23.64	18.18	46.76
CURV_stage2	10.18	5.32	2.75	1.27	18.64	13.93	44.46
CURV	**29.23**	**18.76**	**12.08**	**6.86**	**38.30**	**39.08**	**54.89**

Table 4: Clinical accuracy metrics for radiology report generation on IU X-ray dataset

Model	**CheXbert**			**RadGraph**	
	Acc.	**Macro F1**	**Micro F1**	**Ent. F1**	**F1**
LLaVA-1.5-7B	72.03	4.66	46.81	13.37	8.76
LLaVA-1.5-7B-SFT-CXR	76.34	5.84	53.33	16.19	10.31
HuatuoGPT-Vision-7B	89.89	5.72	67.07	22.98	13.96
MAIRA-2	88.74	**6.22**	70.75	34.53	24.01
Qwen2.5-VL-3B	80.64	4.92	49.47	12.70	6.26
CURV_stage1	68.89	4.57	40.30	20.59	12.13
CURV_stage2	67.59	3.76	33.67	12.83	7.28
CURV	**91.56**	5.86	**74.36**	**36.99**	**25.65**

Out-of-Distribution Evaluation To assess the framework's generalization capabilities, we conducted a rigorous out-of-distribution (OOD) evaluation using the IU X-ray [9] dataset, which was not used during training. As shown in the results (Tables 3 and Tables 4), CURV maintains its strong performance, outperforming all baseline methods on this new dataset. This successful performance demonstrates that the model's learned capabilities for reasoning and expressing uncertainty are robust and can generalize well beyond the MIMIC-CXR dataset it was trained on. These findings provide strong evidence that CURV is not overfit to its training data and can be effectively applied to different clinical data sources.

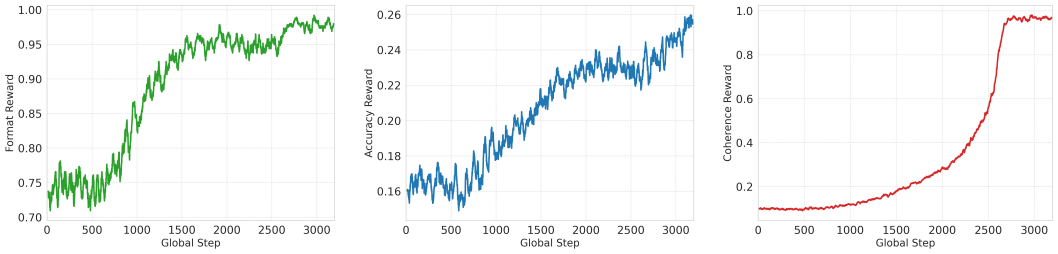
LLM-based Evaluation The LLM-based evaluation, as shown in Figure 3, demonstrates substantial qualitative improvements in CURV-generated reports throughout the reinforcement learning phase. This evaluation is based on criteria defined in Appendix C. The "Thinking" section shows consistent and significant enhancements across all assessed criteria—*Logical Coherence*, *Depth of Analysis*, *Relevance*, *Evidence-Based Nature*, and *Consistency*—with the corresponding scores progressively rising (e.g., *Logical Coherence* from 2.23 to 4.55 and *Consistency* from 2.88 to 4.76 by step 3,200).



(a) LLM-based thinking evaluation (b) LLM-based uncertainty evaluation

Figure 3: LLM-based evaluation of report quality during reinforcement learning. (a) shows the progressive improvement of the 'Thinking' section across five qualitative criteria. (b) tracks the increasing scores for both structural and semantic uncertainty expression, demonstrating the model's refinement over training steps.

This indicates the model's increasing ability to articulate a transparent and sound reasoning process. Regarding uncertainty expression, Semantic Uncertainty scores steadily improved from 1.60 to 2.60, indicating better conveyance of overall diagnostic confidence. Structural Uncertainty scores also increased from an initial value of 0.60 to 0.76 (following an early dip), signifying progress in articulating confidence for specific findings, albeit with more complex learning dynamics observed. Collectively, these trends underscore the efficacy of the CURV framework, especially its RL stage, in fostering clinically nuanced reports that are more interpretable, trustworthy, and adept at expressing diagnostic uncertainty, thereby enhancing clinical utility.

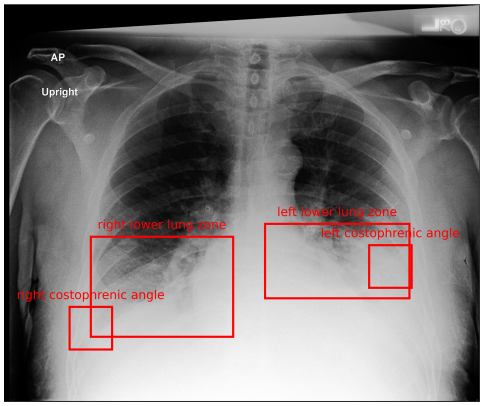


(a) Format Reward (b) Accuracy Reward (c) Coherence Reward

Figure 4: Reward trends during the reinforcement learning phase. The plots show the rapid convergence of the (a) Format Reward, the steady increase of the (b) Accuracy Reward, and the significant late-stage ascent of the (c) Coherence Reward, validating the effectiveness of the multi-component reward function.

Reward Changing The evolution of reward components during reinforcement learning, as depicted in Figure 4, underscores the efficacy of the CURV training strategy. The Format Adherence Reward (R_{fmt}) exhibits a rapid increase early in training, quickly reaching near-optimal values (Figure 4a). This indicates the model's swift adoption of the required tripartite report structure. Concurrently, the Accuracy Reward (R_{acc}), encompassing both entity matching and uncertainty alignment, demonstrates a steady, more gradual improvement throughout the training process (Figure 4b). This reflects the nuanced challenge of enhancing medical accuracy and appropriate uncertainty expression in the "Findings" and "Impression" sections. Most notably, the Thinking Coherence Reward (R_{coh}) initially remains low but undergoes a significant and steep ascent in later training stages, eventually plateauing at a high level (Figure 4c). This trajectory strongly suggests that the reinforcement learning phase, guided by R_{coh}, successfully teaches the model to generate a logically sound "thinking" process that effectively connects radiological findings to clinical impressions. Collectively, these trends validate the multi-component reward function and the RL approach in progressively refining the model towards generating structurally correct, clinically accurate, and coherently reasoned radiology reports.

Case Study To qualitatively illustrate the model's advanced capabilities, we present an analysis of Case Study in Figure 5. The model proficiently generated a clinically coherent and structured report, comprising distinct "Findings," "Thinking," and "Impression" sections, adhering to the desired output



(a) Chest X-ray image, the "bibasal atelectasis" mentioned in both findings and impression is corresponding to right and left lower lung zone as annotated in the image, and the "pleural effusions" is related with left and right costophrenic angle.

Strong Correspondence in Key Findings:

Generated Findings: *"...bibasal atelectasis... There is likely bilateral pleural effusions of moderate extent."*

Ground Truth Findings: *"GT: ...bibasilar opacities compatible with... adjacent atelectasis. Persistent moderate bilateral pleural effusions..."*

Alignment in Overall Clinical Assessment:

Generated Impression: *"...bibasal atelectasis and moderate pleural effusions... the concern is **likely** the effects of... management of any underlying atelectasis and effusions..."*

Ground Truth Impression: *"GT: Persistent moderate bilateral pleural effusions with adjacent atelectasis."*

(b) Generated report sections.

Figure 5: Case Study: Illustrating robust VLM performance in identifying key clinical findings and generating a useful reasoning process. (b) Generated report sections demonstrate strong correspondence with ground truth on core observations such as bibasilar atelectasis and moderate pleural effusions, including appropriate use of uncertainty terms (e.g., "likely"). The model's explicit "<thinking>" section (analyzed in the main text, full content at D) provides a valuable, transparent pathway by linking these findings to patient context (e.g., supine position, HF history) and potential implications, showcasing the utility of structured reasoning in enhancing clinical report generation.

format. Due to page limit, key excerpts are shown, the full case is detailed in Appendix D. In the Findings section, the model accurately identified significant pathologies, including bibasilar atelectasis and moderate bilateral pleural effusions. These observations demonstrated strong correspondence with the ground truth report, and the model appropriately applied structural uncertainty terms (e.g., "likely"), showcasing its uncertainty-aware learning. The "Thinking" section proved particularly valuable, offering a transparent reasoning pathway. It successfully linked the identified visual findings with crucial patient context, such as supine positioning and history of biventricular heart failure, and considered their clinical implications. This explicit articulation of the thought process significantly enhances report interpretability and trustworthiness. Finally, the "Impression" section aligned well with the ground truth's primary assessment, correctly summarizing the key findings of atelectasis and effusions and focusing on their management. This case exemplifies the model's robust ability to produce accurate, contextually-aware, and clinically useful radiological interpretations with a clear and valuable reasoning structure, highlighting the strengths of our proposed framework.

5 Conclusion

We introduced CURV, a novel framework that enhances radiology report generation by integrating uncertainty awareness and explicit reasoning into vision-language models. CURV's key innovations—uncertainty modeling, a structured reasoning module, and a coherence-driven reinforcement learning strategy—enable the generation of reports that are clinically accurate, transparent, and appropriately express diagnostic confidence. Experimental results demonstrate CURV's superior performance in producing interpretable AI-generated CXR reports. However, CURV's performance relies on the quality of the initial curated datasets for uncertainty and reasoning, and its "thinking" process, while explicit, is learned via LLM-generated data. Furthermore, its generalization to other medical imaging modalities requires further investigation. Ultimately, the most critical next step is a large-scale clinical validation by expert radiologists to ensure its safe and effective translation into practice. Despite these limitations, CURV marks a significant advancement for trustworthy vision-language models in high-stakes clinical applications. To further promote research and transparency in this domain, the TRACE-CXR dataset developed for this study will be made publicly available. We believe CURV and the TRACE-CXR dataset will serve as valuable resources for future work where clarity in reasoning and uncertainty is crucial.

Acknowledgments and Disclosure of Funding

This work is partially supported by the National Natural Science Foundation of China (62302413), the Health and Medical Research Fund (23220312), the General Research Fund RGC/HKBU12202621 from the Research Grant Council, and the Research Matching Grant Scheme RMGS2021_8_06 from the Hong Kong Government.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

[3] Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation, 2024. URL https://arxiv.org/abs/2406.04449.

[4] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. Sft or rl? an early investigation into training r1-like reasoning large vision-language models, 2025. URL https://arxiv.org/abs/2504.11468.

[5] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, 2024. URL https://arxiv.org/abs/2406.19280.

[6] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=dYur3yabMj.

[7] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[8] Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.319. URL https://aclanthology.org/2022.findings-emnlp.319/.

[9] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza M. Rodriguez, Sameer Kiran 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 : JAMIA*, 23 2:304–10, 2015. URL https://api.semanticscholar.org/CorpusID:16941525.

[10] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, ..., and H. E. Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation [Online]*, 101(23):e215–e220, 2000.

[11] Difei Gu, Yunhe Gao, Yang Zhou, Mu Zhou, and Dimitris Metaxas. Radalign: Advancing radiology report generation with vision-language concept alignment, 2025. URL https://arxiv.org/abs/2501.07525.

[12] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models, 2025. URL https://arxiv.org/abs/2503.06749.

[13] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. Nguyen Duong, T. Bui, P. Chambon, M. Lungren, A. Ng, C. Langlotz, and P. Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports (version 1.0.0), 2021. URL https://doi.org/10.13026/hm87-5p47.

[14] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=pMWtc5NKd7V.

[15] Peiyuan Jing, Kinhei Lee, Zhenxuan Zhang, Huichi Zhou, Zhengqing Yuan, Zhifan Gao, Lei Zhu, Giorgos Papanastasiou, Yingying Fang, and Guang Yang. Reason like a radiologist: Chain-of-thought and reinforcement learning for verifiable report generation, 2025. URL https://arxiv.org/abs/2504.18453.

[16] A. Johnson, T. Pollard, R. Mark, S. Berkowitz, and S. Horng. Mimic-cxr database (version 2.1.0), 2024. URL https://doi.org/10.13026/4jqj-jw95.

[17] A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, and et al. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data*, 6(317), 2019. doi: 10.1038/s41597-019-0322-0. URL https://doi.org/10.1038/s41597-019-0322-0.

[18] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[19] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

[20] Chunyu Liu, Yixiao Jin, Zhouyu Guan, Tingyao Li, Yiming Qin, Bo Qian, Zehua Jiang, Yilan Wu, Xiangning Wang, Ying Feng Zheng, and Dian Zeng. Visual–language foundation models in medicine. *The Visual Computer*, 41(4):2953–2972, March 2025. ISSN 1432-2315. doi: 10.1007/s00371-024-03579-w. URL https://doi.org/10.1007/s00371-024-03579-w.

[21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296, 2024. doi: 10.1109/CVPR52733.2024.02484.

[22] Ivona Najdenkoska, Xiantong Zhen, Marcel Worring, and Ling Shao. Uncertainty-aware report generation for chest x-rays by variational topic inference. *Medical Image Analysis*, 82:102603, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102603. URL https://www.sciencedirect.com/science/article/pii/S1361841522002341.

[23] Toru Nishino, Yasuhide Miura, Tomoki Taniguchi, Tomoko Ohkuma, Yuki Suzuki, Shoji Kido, and Noriyuki Tomiyama. Factual accuracy is not enough: Planning consistent description order for radiology report generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7123–7138, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.480. URL https://aclanthology.org/2022.emnlp-main.480/.

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

[25] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, and et. al. Capabilities of Gemini Models in Medicine, May 2024. URL http://arxiv.org/abs/2404.18416.

[26] Luca M. Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7(1):96–106, January 2025. ISSN 2522-5839. doi: 10.1038/s42256-024-00963-y. URL https://doi.org/10.1038/s42256-024-00963-y.

[27] Yongqi Shao, Renxin Xu, Cong Tan, Gaofeng Liu, Tao Fang, and Hong Huo. Reinforcement learning for improved alignment in radiology reports generation. In *2024 China Automation Congress (CAC)*, pages 6250–6255, 2024. doi: 10.1109/CAC63892.2024.10864815.

[28] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

[29] Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated Radiology Report Generation: A Review of Recent Advances. *IEEE reviews in biomedical engineering*, 18:368–387, 2025. ISSN 1941-1189. doi: 10.1109/RBME.2024.3408456.

[30] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117/.

[31] Ryutaro Tanno, David G. T. Barrett, Andrew Sellergren, Sumedh Ghaisas, and et. al. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31 (2):599–608, February 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03302-1. URL https://doi.org/10.1038/s41591-024-03302-1.

[32] Yixin Wang, Zihao Lin, Zhe Xu, Haoyu Dong, Jie Luo, Jiang Tian, Zhongchao Shi, Lifu Huang, Yang Zhang, Jianping Fan, and Zhiqiang He. Trust it or not: Confidence-guided automatic radiology report generation. *Neurocomputing*, 578:127374, April 2024. ISSN 09252312. doi: 10.1016/j.neucom.2024.127374. URL https://linkinghub.elsevier.com/retrieve/pii/S0925231224001450.

[33] J. Wu, N. Agu, I. Lourentzou, A. Sharma, J. Paguio, J. S. Yao, E. C. Dee, W. Mitchell, S. Kashyap, A. Giovannini, L. A. Celi, T. Syeda-Mahmood, and M. Moradi. Chest imagenome dataset (version 1.0.0), 2021. URL https://doi.org/10.13026/wv01-y230.

[34] Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, Honglin Guo, Wei Shen, Xiaoran Fan, Yuhao Zhou, Shihan Dou, Xiao Wang, Xinbo Zhang, Peng Sun, Tao Gui, Qi Zhang, and Xuanjing Huang. Training large language models for reasoning through reverse curriculum reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[35] Sixing Yan, Haiyan Yin, Ivor W. Tsang, and William K. Cheung. Diagnose with uncertainty awareness: Diagnostic uncertainty encoding framework for radiology report generation. In Carole H. Sudre, Raghav Mehta, Cheng Ouyang, Chen Qin, Marianne Rakic, and William M. Wells, editors, *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 34–44, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73158-7.

[36] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and et. al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[37] Hao-Yu Yang, Junling Yang, Yue Pan, Kunlin Cao, Qi Song, Feng Gao, and Youbing Yin. Learn to be uncertain: Leveraging uncertain labels in chest x-rays with bayesian neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[38] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S. Kevin Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86: 102798, 2023. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.102798. URL https://www.sciencedirect.com/science/article/pii/S1361841523000592.

[39] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[40] Yuexiang Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Shengbang Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. Fine-tuning large vision-language models as decision-making agents via reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=nBjmMF2IZU.

[41] Shuang Zhou, Jiashuo Wang, Zidu Xu, Song Wang, David Brauer, Lindsay Welton, Jacob Cogan, Yuen-Hei Chung, Lei Tian, Zaifu Zhan, et al. Uncertainty-aware large language models for explainable disease diagnosis. *arXiv preprint arXiv:2505.03467*, 2025.

[42] Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. Large model driven radiology report generation with clinical quality reinforcement learning. *CoRR*, abs/2403.06728, 2024. URL https://doi.org/10.48550/arXiv.2403.06728.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 1

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper proposes a novel framework and methodology (CURV) for X-Ray report generation, focusing on its empirical performance and practical contributions. While it uses mathematical formulations to define its components, loss functions, and reward mechanisms, it does not present new theoretical results such as theorems or lemmas accompanied by formal assumptions and proofs. The primary validation is experimental.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Section 5 states the TRACE-CXR dataset will be made publicly available. The core dataset, MIMIC-CXR, is also accessible. Code will be made available once accepted.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: While statistical validation through error bars or significance testing is valuable, it was not included for the main quantitative results in this submission. This paper introduces CURV, a novel and complex multi-stage framework for a challenging task. The primary focus of this foundational study was to establish the framework's architecture, demonstrate its unique capabilities in uncertainty modeling and explicit reasoning, and show its potential to significantly advance performance on key clinical metrics. Given the extensive computational resources and time required to conduct multiple full end-to-end training runs of such a complex system, a comprehensive statistical analysis was beyond the scope of this initial investigation. We acknowledge this as a limitation and plan to incorporate rigorous statistical validation in future in-depth studies and extensions of this work.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4, Appendix A

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed and conform to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Section 1 details the positive societal benefits. Potential negative impacts, such as over-reliance or algorithmic bias common to AI in healthcare, are acknowledged as areas requiring ongoing attention in the field, though not extensively detailed for CURV specifically.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: The paper announces the future public release of the TRACE-CXR dataset. Specific safeguards for this release (e.g., data use agreements, specific licensing terms to prevent misuse) are not detailed in this version but will be considered upon release. The underlying MIMIC-CXR data is accessed under its established protocols.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Appendix E

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Section 4

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No crowdsourcing used

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: Does not involve crowdsourcing nor research with human subjects

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper describes the usage of LLMs as they form an important and, in part, original component of the core methodology, particularly in dataset creation/augmentation and evaluation. Specifically, LLMs (Qwen2.5-7B-instruct, grok-3) were utilized to generate and augment critical training data. This includes generating "thinking" sections for our novel **TRACE-CXR** dataset, a contribution of this work used for reasoning initialization (detailed in Section 1; Section 3, subsection on "Reasoning Initialization for Transparent Reporting"; and Appendix B). LLMs also derived uncertainty annotations for the uncertainty modeling mechanism (see Section 3, subsection on "Empowering Uncertainty Awareness in Vision-Language Models"; and Appendix B), and enhanced parsing of the MIMIC-CXR dataset (Appendix B). These LLM-generated/augmented datasets are fundamental to the supervised fine-tuning (SFT) stages of the CURV framework. Furthermore, an LLM (Qwen3-32B) was employed for parts of our evaluation protocol, specifically for the qualitative assessment of the generated "Thinking" sections and uncertainty expressions, a non-standard approach (detailed in Subsection 4.3 and Appendix C). This overall usage is detailed appropriately within the manuscript as it is integral to our data preparation, model training methodology, evaluation strategy, and constitutes some of the original contributions of the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Experimental Setup

In this section, we outline the experimental setup for training and evaluating the CURV framework for radiology report generation and the comparison methods.

Training Configuration Experiments were conducted on a node with 4xA100-80G GPUs. We utilized the Qwen-2.5-VL-3B as the backbone model for its balance of efficiency and performance. The training was configured with a batch size of 16, and a group size of 8 for GRPO in the reinforcement learning stage. The entire process spanned 3,200 steps over approximately 100 hours, with a learning rate of 1×10^{-6} used for stable convergence.

To manage the memory requirements of training a 3B parameter model, we employed different fine-tuning strategies and state-of-the-art memory optimization techniques across the stages:

- **Stage 1 and 2 (SFT - Uncertainty Aware and Reasoning Initialization)**: To maximize efficiency, we kept the vision encoder frozen and fine-tuned only the parameters of the multimodal projection layer and the LLM backbone. This stage took approximately 4 hours to complete on 4 GPUs.
- **Stage 3 (RL - Enhancement)**: In this final stage, the goal was to refine the entire model's behavior based on global reward signals. Therefore, we performed full-parameter fine-tuning of the model checkpoint from Stage 2. To make this computationally intensive step feasible within the 80GB memory of each GPU, we leveraged a combination of standard optimization techniques:
  - **Mixed-Precision Training**: We used bfloat16 (bf16) precision, which halves the memory footprint for model parameters, gradients, and optimizer states compared to standard FP32, with minimal impact on training stability.
  - **Parameter and Optimizer State Sharding**: We utilized DeepSpeed with ZeRO Stage 3 optimization. This powerful technique partitions not only the optimizer states but also the gradients and the model parameters themselves across all 4 GPUs. This drastically reduces the per-GPU memory load, as each GPU only holds a fraction of the total training-related tensors.
  - **Activation Checkpointing**: To manage memory consumed by activations, we employed activation checkpointing (also known as gradient checkpointing). This technique avoids storing all activations by re-computing them during the backward pass, trading a modest amount of compute time for a significant reduction in memory usage.

Baseline Models For the baseline models, CURV was benchmarked against established vision-language models, briefly introduced below with their experimental configurations:

- **LLaVA-1.5-7B** [21]: A vision language model for image and language understanding, fine-tuned on Vicuna with GPT-generated data for strong chat capabilities. We used the original model and a variant, LLaVA-1.5-7B-SFT-CXR, fine-tuned on a chest X-ray (CXR) dataset for 1 epoch.
- **MAIRA-2** [3]: A vision language model for radiology report generation from chest X-rays, we used the original model without additional fine-tuning.
- **HuatuoGPT-Vision-7B** [5]: A medical vision language model based on Qwen2-7B and LLaVA-v1.5, trained on PubMedVision for medical vision-language tasks. The original model was used without further fine-tuning.

All baselines were evaluated under consistent conditions, aligning with dataset and metric details in Sections 4.2 and 4.3, to ensure a fair comparison with CURV for chest X-ray report generation.

B Datasets

We details the datasets and preprocessing steps undertaken to train and evaluate the CURV framework. A key contribution of our work is the curation and enhancement of existing datasets to specifically support uncertainty modeling and structured reasoning in radiology report generation.

Table 5: Dataset Statistics for CURV Training and Evaluation

Statistic	Value
MIMIC-CXR (Original)	
Total Reports	227,835
Reports with Findings (REGEX)	65.7%
Reports with Impression (REGEX)	82.3%
Reports with Both Findings & Impression (REGEX)	48.0%
MIMIC-CXR (After LLM-Enhanced Parsing)	
Reports with Findings	185,122 (81.25%)
Reports with Impression	193,755 (85.04%)
Reports with Both Findings & Impression	151,048 (66.30%)
Uncertainty Dataset ($\mathcal{D}_{\text{uncertainty}}$)	
Samples with Uncertainty Annotations	112,111
Unique Uncertainty Expressions (freq. > 5)	$\sim$2,700
TRACE-CXR Dataset ($\mathcal{D}_{\text{reason}}$)	
Number of Reports in **TRACE-CXR**	2000

Core Dataset and LLM-Enhanced Parsing Our primary dataset is MIMIC-CXR [16, 17, 10], a large-scale collection of 227,835 radiology reports. Initial analysis using REGEX-based parsing revealed considerable heterogeneity in report structure. While "IMPRESSION" sections were present in 82.3% of reports and "FINDINGS" in 65.7%, only 48.0% of reports reliably contained both. This structural variability, coupled with an average raw text length of 634.3 characters (Findings: 335.2 chars, Impression: 175.1 chars), hindered consistent extraction of these crucial sections. To address these limitations and create a more uniform dataset for subsequent annotation and training, we employed a Large Language Model (LLM) for enhanced parsing of the MIMIC-CXR reports. This step significantly improved the availability of structured data: the proportion of reports with an identifiable "FINDINGS" section increased to 81.25% (185,122 reports), and those with an "IMPRESSION" section rose to 85.04% (193,755 reports). Consequently, the number of reports containing both "FINDINGS" and "IMPRESSION" sections increased substantially from 48.0% to 66.30% (151,048 reports), providing a more robust foundation for our work.

Curating Data for Uncertainty Modeling A core innovation of CURV is its explicit modeling of diagnostic uncertainty. To train this capability (Stage 1 of our pipeline), we created a specialized uncertainty-annotated dataset, $D_{uncertainty}$. This was achieved by leveraging the Imagenome dataset [33], which provides valuable links between textual phrases in radiology reports and corresponding bounding box localizations on CXR images. Building upon these existing spatial annotations from Imagenome, we utilized the Qwen2.5-7B-instruct model to perform fine-grained uncertainty extraction. This model was specifically prompted to identify and extract uncertainty-expressing phrases directly from the report sentences that Imagenome had linked to specific visual findings. This process allowed us to map linguistic expressions of uncertainty to precise image regions. Our final uncertainty-annotated dataset contains 112,111 samples. Through this process, we identified approximately 2,700 unique uncertainty expressions that occurred with a frequency greater than five. The most common expressions included "likely" (131,334 instances), "may" (91,206 instances), and "could" (71,267 instances). For training, these annotations were formatted into structured JSON outputs, each containing the bounding box coordinates, an anatomical label, and the specific uncertainty description related to that finding (e.g., `{"bbox_2d": [121, 104, 180, 162], "label": "left lower lung zone", "uncertainty":"Bilateral nodular opacities that most likely represent nipple shadows."}`).

Generating the TRACE-CXR Dataset ($\mathcal{D}_{\text{reason}}$) for Reasoning Initialization To initialize the basic reasoning capabilities in Stage 2 of the CURV framework, we constructed the **TRACE-CXR** dataset (also referred to as $\mathcal{D}_{\text{reason}}$ in our methodological descriptions in Section 3). The goal was to create training instances that explicitly model a "thinking" process connecting radiological findings to clinical impressions. The exact prompt used to guide the LLM in emulating an experienced radiologist's logical, step-by-step reasoning process is shown in Figure 6. Additionally, we validated the generated data using a separate prompt to ensure quality, retaining only those entries with a score higher than 80 out of 100, as shown in 7. Using the LLM-enhanced "FINDINGS" and "IMPRES-

```
You are a highly experienced radiologist tasked with generating a
    detailed reasoning process that explains how specific findings in
    a radiology report lead to a clinical impression. Your goal is to
    create a logical, step-by-step explanation that connects the
    observed features in a medical image to the final diagnosis or
    clinical takeaway. Use precise medical terminology and ensure the
    reasoning is clear, concise, and relevant to the provided
    findings and impression.
Input:
- Findings: [content of findings]
- Impression: [content of impression]
Task:
1. Analyze the provided findings and impression.
2. Generate a detailed reasoning process that explains how the
    findings support the impression. Break down the explanation into
    logical steps, addressing:
   - What specific abnormalities or features in the findings are most
      relevant to the impression.
   - How these features are typically associated with the suspected
      condition or diagnosis.
   - Any differential diagnoses or alternative possibilities
      considered based on the findings, and why the given impression
      is the most likely.
   - If applicable, mention any additional context (e.g., typical
      clinical presentations, risk factors, or imaging
      characteristics) that supports the reasoning.
3. Enclose your reasoning process in <thinking> tags as follows:
    <thinking>Your detailed reasoning here</thinking>.
Output Format:
<thinking>
[Your detailed step-by-step reasoning connecting the findings to the
    impression]
</thinking>
```

Figure 6: The LLM prompt for generating the 'thinking' section in the TRACE-CXR dataset. This prompt guides the LLM to create a logical, step-by-step reasoning process that connects the provided findings to the clinical impression, emulating a radiologist's thought process.

SION" sections as inputs, we prompted an LLM (grok-3) to generate an intermediate "THINKING" section. The LLM was guided by detailed instructions to emulate an experienced radiologist, tasking it to analyze the provided findings and impression, and then to construct a logical, step-by-step explanation of how the findings support the impression, including consideration of differential diagnoses where appropriate. The resulting data for $\mathcal{D}_{\text{reason}}$ consists of the CXR image paired with a structured report containing three distinct sections: `<findings>Detailed description of image observations</findings>`, `<thinking>Reasoning based on findings</thinking>`, and `<impression>Concise summary and recommendations</impression>`. Table 5 summarizes key statistics of the datasets pivotal to CURV's training and evaluation.

Table 6: Comprehensive comparison of evaluation scores between Clinicians (Clin), Grok3 (Grok), and Gemini 2.5 Pro (Gem).

Study ID	Logical Coherence			Depth of Analysis			Relevance			Evidence Based			Consistency			Overall Score	
	Clin	Grok	Gem	Clin	Grok	Gem	Clin	Grok	Gem	Clin	Grok	Gem	Clin	Grok	Gem	Grok	Gem
s54517467	4	5	5	5	5	5	5	5	5	4	5	5	4	5	5	5.0	5.0
s51966501	4	5	5	5	5	5	4	5	5	4	5	5	4	5	5	5.0	5.0
s52188295	5	5	5	5	5	5	4	5	5	4	5	5	5	5	5	5.0	5.0
s55493024	4	5	5	5	5	5	4	5	5	4	5	5	5	5	5	5.0	5.0
s51074196	4	5	5	5	5	5	4	5	5	4	5	5	5	5	5	5.0	5.0
s56689492	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5.0	5.0
s57002637	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5.0	5.0
s50849849	5	5	5	5	5	5	5	4	5	5	4	5	5	3	4	4.2	4.8
s51215354	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5.0	5.0
s52528325	4	5	5	5	5	5	4	5	5	4	5	5	5	5	5	5.0	5.0

Clinical Expert Evaluation of TRACE-CXR Dataset To validate the quality of the LLM-generated TRACE-CXR dataset and the reliability of our LLM-based evaluation protocol, we conducted a formal evaluation with a board-certified radiologist. We compared their expert judgments against those of two state-of-the-art LLMs (Grok3 and Gemini 2.5 Pro). For the evaluation, a random sample of 10 reports from the TRACE-CXR dataset was scored by all three evaluators using the exact same criteria outlined in Appendix C (Logical Coherence, Consistency, etc.). The results, summarized in Table 6, reveal a strong alignment between the human expert and both LLMs. Specifically, Gemini 2.5 Pro's scores were within one point of the clinician's 100% of the time, with a very low Mean Absolute Error (MAE) of 0.380. Similarly, Grok3's scores were within one point of the expert's 98.0% of the time, with an MAE of 0.440. This high degree of concordance provides strong, multi-faceted validation for our approach. It not only confirms that the LLM-generated reasoning traces in the TRACE-CXR dataset are of high clinical quality but also substantiates that our LLM-based evaluation metrics serve as a reliable proxy for expert human judgment.

```
**Task:** Evaluate a chest X-ray (CXR) report's <thinking> section
    across five key aspects.
**Instruction:** Assess the <thinking> section of a chest X-ray
    report across multiple aspects. Analyze how well the reasoning
    links findings to impressions, its depth, clinical relevance,
    evidence base, and consistency with other sections.
**Input Format:**
- `<findings>`: Observations from the CXR image.
- `<thinking>`: Reasoning or analysis based on findings.
- `<impression>`: Summarized conclusions or diagnosis.
**Provided Input:**
{findings}{thinking}{impression}
**Evaluation Criteria:**
Evaluate ALL of the following aspects independently:
1. **Logical Coherence:**
   - **Score 5:** Clear, logical flow, seamlessly connecting findings
       to impressions without doubt.
   - **Score 0:** Incoherent or illogical; reasoning is fragmented or
       fails to connect findings to impressions.
2. **Depth of Analysis:**
   - **Score 5:** Deep analysis with comprehensive explanations,
       including alternatives or limitations.
   - **Score 0:** Superficial or absent analysis; merely restates
       findings without insight.
3. **Relevance:**
   - **Score 5:** Highly relevant, focusing on key clinical findings
       without extraneous content.
   - **Score 0:** Irrelevant or off-topic reasoning, ignoring
       clinical context.
4. **Evidence-based:**
   - **Score 5:** Strongly evidence-based, tied to medical knowledge
       and practices.
   - **Score 0:** Lacking evidence; speculative or contrary to
       medical standards.
5. **Consistency:**
   - **Score 5:** Fully consistent across `<findings>`, `<thinking>`,
       and `<impression>`.
   - **Score 0:** Grossly inconsistent with contradictions
       undermining trustworthiness.
**Output Format:** Return evaluation as a JSON object per the
    provided schema.
```

Figure 7: The LLM-based evaluation prompt for the 'Thinking' section. This prompt instructs the evaluator LLM to score the generated reasoning on five criteria—Logical Coherence, Depth of Analysis, Relevance, Evidence-Based Nature, and Consistency—and return the assessment in a structured JSON format.

```
**Task:** Comprehensively evaluate uncertainty expressions in a chest
    X-ray (CXR) report, comparing a generated report with a ground
    truth report.
**Instruction:** You are tasked with rigorously assessing how well
    the generated report expresses uncertainty compared to the ground
    truth report. Focus on two key aspects:
1. **Structural Uncertainty:** Hedging or ambiguity about specific
    anatomical regions or findings (typically in findings sections)
   - Example: "a nodule may represent a benign lesion or malignancy"
   - This appears in descriptions of specific observations
2. **Semantic Uncertainty:** Ambiguity in overall diagnostic
    synthesis (typically in impression sections)
   - Example: "findings are nonspecific and could be consistent with
       infection"
   - This appears in the overall assessment/conclusion
First determine which parts of each report represent findings vs.
    impression sections, even if they aren't explicitly labeled. Then
    compare the uncertainty expressions between generated and ground
    truth reports.
**Evaluation Criteria:**
1. **Structural Uncertainty:**
   - **Score 5:** Generated report contains uncertainty expressions
       about specific findings highly similar to ground truth in both
       content and strength
   - **Score 3:** Somewhat similar with noticeable differences
   - **Score 0:** No structural uncertainty or completely dissimilar
       to ground truth
2. **Semantic Uncertainty:**
   - **Score 5:** Generated report contains holistic diagnostic
       uncertainty expressions highly similar to ground truth
   - **Score 3:** Somewhat similar with noticeable differences
   - **Score 0:** No semantic uncertainty or completely dissimilar to
       ground truth
**Output Format:** Return your evaluation as a JSON object with two
    main sections (structural_uncertainty and semantic_uncertainty)
    each containing:
- score (0-5)
- explanation (1-2 sentences justifying the score)
- triples_comparison (comparison of uncertainty triples - subject,
    uncertainty term, interpretation)
- uncertainty_strength (comparison of strength of uncertainty terms)
- contextual_appropriateness (assessment of whether uncertainty is
    expressed in appropriate contexts)
Also include an overall_score section with the same fields (calculate
    score as average of the two aspects).
**Provided Input:**
Generated Report:
{generated_text}
Ground Truth Report:
{ground_truth_text}
```

Figure 8: The LLM-based evaluation prompt for uncertainty expression. This prompt directs the evaluator LLM to assess and score the generated report's handling of both structural (finding-specific) and semantic (diagnostic) uncertainty in comparison to the ground truth report.

C Evaluation Metric

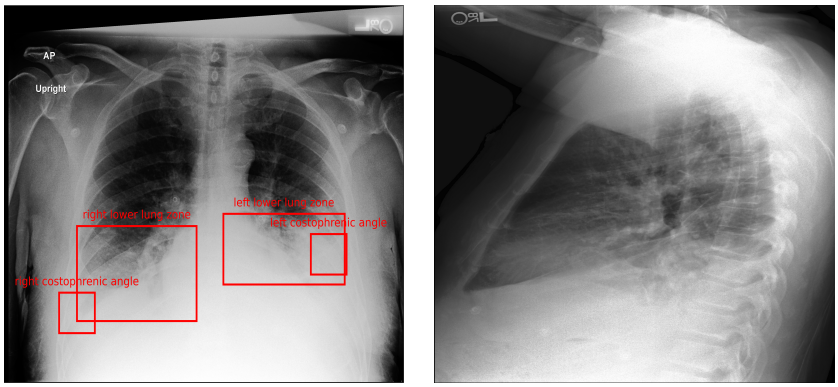
To comprehensively assess the performance of CURV and baseline methods, we employ a suite of evaluation metrics targeting clinical accuracy, the quality of uncertainty expression, and the coherence of the generated reasoning pathways.

Clinical Accuracy and NLP Metrics For evaluating the factual correctness and fluency of the generated "Findings" and "Impression" sections, we utilize standard Natural Language Processing (NLP) metrics, including BLEU (1-4) [24], ROUGE-L [19], METEOR [2]. Beyond these, we measure clinical accuracy using F1-scores based on medical entity extraction. Specifically, we employ CheXbert [30] and Radgraph [13] to calculate the f1 score of extract entities.

LLM-based Evaluation To evaluate the "Thinking" section generated by CURV, which is not present in standard reports and thus not amenable to traditional reference-based metrics, we leverage an LLM-based evaluation protocol. We prompt a capable LLM (Qwen3-32B [36]) to assess the quality of the "Thinking" section based on five crucial aspects: Logical Coherence, Depth of Analysis, Relevance, Evidence-Based Nature, and Consistency with the "Findings" and "Impression" sections. And we also evaluate the expression of uncertainty, particularly distinguishing between structural and semantic uncertainty, also requires a more sophisticated approach than standard metrics. We employ another LLM-based evaluation protocol for this purpose. The LLM is tasked with comparing the uncertainty expressions in a generated report against those in a ground truth report, focusing on two key aspects: Structural Uncertainty and Semantic Uncertainty. The prompt for the LLM-based evaluation is shown in Figure 7 and 8.

D Full Case Study

This appendix provides the complete details to case study as shown in Figure 9.



(a) Frontal and Lateral Images

Ground Truth: Findings and Impression

FINDINGS: There are persistent bibasilar opacities compatible with layering effusions with adjacent atelectasis and perhaps mild pulmonary edema. The cardiomediastinal silhouette is stable. Atherosclerotic calcifications seen at the aortic arch. No acute osseous abnormalities.
IMPRESSION: Persistent moderate bilateral pleural effusions with adjacent atelectasis. Superimposed infection would be difficult to exclude.

Model Output: Findings, Thinking and Impression

FINDINGS: He has bibasal atelectasis, likely related to poor mobilization of lung volumes during inspiration. There is likely bilateral pleural effusions of moderate extent. However, there are no pulmonary parenchymal opacities at the apices; the pleural surface appears clear. Given patient's history of biventricular HF, she is supine on the exam, resulting in her moderate and greater than moderate apical lung collapse. No evidence of new cardiomegaly or acute pulmonary vascular congestion.
THINKING: Since the imaging shows bibasal atelectasis, a clear pleural surface, and no new signs of cardiomegaly or pulmonary vascular congestion, the history of biventricular heart failure likely plays a significant role. The patient is described as supine on the exam, which contributes to the bibasal atelectasis but does not suggest acute worsening of the pulmonary condition. The presence of moderate pleural effusions is noted, and their lack of apical opacities could indicate no underlying pulmonary parenchymal disease such as pneumonia. Therefore, the primary concern is likely the presence and management of the atelectasis and effusions, given the patient's cardiopulmonary status.
IMPRESSION: The patient is supine and therefore has bibasal atelectasis and moderate pleural effusions. There is no evidence of pulmonary parenchymal opacities at the apices, and the pleural surface is clear. The presence of moderate pleural effusions suggests that the lung volumes were better mobilized during expiration. With patient's history of biventricular HF, the concern is likely the effects of the supine positioning on lung volumes and the management of any underlying atelectasis and effusions, though no acute pulmonary or cardiac changes are seen.

Figure 9: A full, detailed view of the case study. This figure presents the frontal and lateral CXR images, the ground truth report, and the complete, three-part report generated by CURV, showcasing the model's ability to produce a clinically coherent and well-structured output.

E Licenses for External Assets

This appendix details the licenses for publicly available datasets and pre-trained models utilized in this work.

Datasets

- **MIMIC-CXR (version 2.1.0):**
  - **Source:** Johnson et al. [17].
  - **URL:** `https://physionet.org/content/mimic-cxr/2.1.0/`
  - **License:** PhysioNet Credentialed Health Data License 1.5.0. Access requires credentialing and signing a data use agreement.

- **Chest ImaGenome Dataset (version 1.0.0):**
  - **Source:** Wu et al. [33].
  - **URL:** `https://physionet.org/content/chest-imagenome/1.0.0/`
  - **License:** Derived from MIMIC-CXR, subject to the PhysioNet Credentialed Health Data License 1.5.0.

Pre-trained Models and Baselines

- **Qwen-2.5-VL-3B (Backbone for CURV):**
  - **Source:** Yang et al. [36].
  - **URL:** `https://github.com/QwenLM/Qwen3`
  - **License:** Apache 2.0 License
- **LLaVA-1.5-7B:**
  - **Source:** Liu et al. [21].
  - **URL:** `https://github.com/haotian-liu/LLaVA`
  - **License:** Apache 2.0 License.
- **MAIRA-2:**
  - **Source:** Bannur et al. [3].
  - **URL:** `https://huggingface.co/microsoft/maira-2`
  - **License:** MICROSOFT RESEARCH LICENSE TERMS
- **HuatuoGPT-Vision-7B:**
  - **Source:** Chen et al. [5].
  - **URL:** `https://huggingface.co/FreedomIntelligence/HuatuoGPT-Vision-7B`
  - **License:** Apache 2.0 License
- **RadGraph:**
  - **Source:** Jain et al. [13].
  - **URL:** `https://huggingface.co/StanfordAIMI/RRG_scorers`
  - **License:** MIT License.
- **CheXbert:**
  - **Source:** Smit et al. [30].
  - **URL:** `https://huggingface.co/StanfordAIMI/RRG_scorers`
  - **License:** MIT License.