UR^2-MLLM: Uncertainty-aware Revisit Reasoning in Multimodal Large Language Models for Radiology Report Generation

Yucheng Chen[1,2,3], Yang Yu[4], Jiazhou Zhou[5], Yufei Shi[1,2], Yongying Lan[1,6], Yichi Zhang[1,2,3], Liyi Li[2], Si Yong Yeo[1,2,3]*

[1]MedVisAI Lab

[2]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

[3]Centre of AI in Medicine, Singapore

[4]Institute of Advanced Intelligence and Computing (IAIC), A*STAR, Singapore

[5]AI Thrust, The Hong Kong University of Science and Technology (Guangzhou)

[6]Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, China

yucheng005@e.ntu.edu.sg siyong.yeo@ntu.edu.sg

Abstract

Radiologists generate diagnostic reports through iterative and selective revisiting of suspicious regions to refine their interpretations. Recent multimodal large language models (MLLMs) for radiology report generation (RRG) have shifted from text-only reasoning toward a "Thinking-with-Images" paradigm, incorporating visual evidence into the reasoning process. However, existing methods provide static visual evidence without a dynamic revisit mechanism during reasoning, neglecting how radiologists re-examine uncertain observations. To this end, we propose an Uncertainty-aware Revisit Reasoning MLLM (UR^2-MLLM) framework that dynamically revisits uncertain regions during reasoning for RRG. UR^2-MLLM is first equipped with uncertainty perception by training on an uncertainty-aware dataset. We then construct a multimodal reasoning trajectory dataset together with a detect-and-copy mechanism, which guides when and where to revisit. Finally, a visual grounding reward refines this behavior through reinforcement learning, aligning the revisited regions with corresponding anatomical structures. Experiments on MIMIC-CXR and IU-Xray show that UR^2-MLLM achieves state-of-the-art performance, highlighting the value of uncertainty-aware visual revisit reasoning for reliable and clinically aligned report generation.

1 Introduction

Radiology report generation (RRG) aims to automatically produce diagnostic narratives from medical images (Chen et al., 2026, 2020), alleviating the heavy workload of radiologists while improving the quality and consistency of clinical documentation (Izhar et al., 2025; Li et al., 2025). In clinical practice, report generation is inherently a decision-making process under uncertain observations, where radiologists repeatedly revisit suspicious regions, cross-reference visual findings with prior knowledge, and progressively refine their interpretations before committing to a final diagnosis (Van der Gijp et al., 2017). Recent advances in multimodal large language models (MLLMs), together with Chain-of-Thought (CoT) (Wei et al., 2022) reasoning techniques, have demonstrated strong potential in approximating this iterative reasoning process, enabling more flexible and clinically aligned report generation.

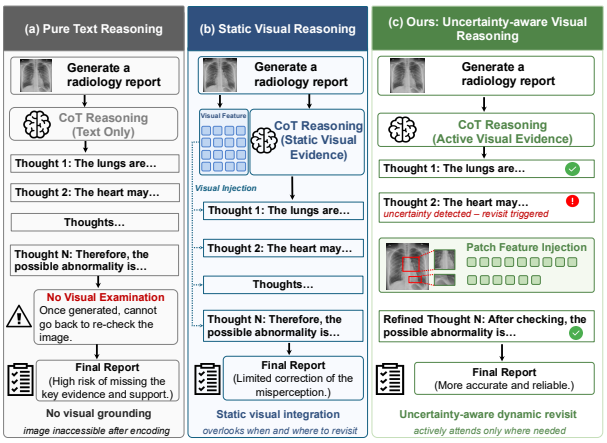


Figure 1: Three reasoning paradigms for radiology report generation. Our UR^2-MLLM triggers uncertainty-aware visual revisits during reasoning.

Existing approaches to CoT-based RRG can be broadly categorized into two paradigms. The first follows a text-only reasoning paradigm (Fig. 1 (a)),

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 14596–14620
October 24-29, 2026 ©2026 Association for Computational Linguistics

where the model encodes the input image once and conducts the entire reasoning process purely in the textual space (Liu et al., 2024b; Ng et al., 2025; Luo et al., 2025; Fan et al., 2025; Wang et al., 2026). As a result, the reasoning trajectory becomes perceptually decoupled from the image, often leading to error accumulation and hallucinated findings (Liu et al., 2025). To mitigate the issue, the second paradigm, often referred to as "Think-with-Images" (Su et al., 2025) (Fig. 1(b)), incorporates static visual evidence throughout the reasoning process. For instance, MAIRA-2 (Bannur et al., 2024) and BoxMed-RL (Jing et al., 2025) provide the whole image as static visual context and output bbox coordinates as text tokens in the reasoning trace. However, this paradigm approximates the diagnostic process with static visual evidence in a single pass, lacking an explicit mechanism to identify and refine uncertain regions as reasoning progresses. Such an approximation implicitly treats perception as deterministic, discarding the diagnostic uncertainty that is critical in clinically ambiguous cases such as overlapping anatomical structures or subtle pathologies. This often leads to over-confident or hallucinated reports, whereas a radiologist would naturally re-examine the relevant regions before finalising the interpretation. Inspired by this, we ask a fundamental research question: ***Can an MLLM learn to revisit visual evidence dynamically, deciding when to revisit based on its own uncertainty and where to revisit based on the corresponding regions?***

To bridge this gap, we propose an Uncertainty-aware Revisit Reasoning MLLM (UR^2-MLLM), a framework that dynamically revisits visual evidence during reasoning, determining *when* and *where* to revisit based on its own uncertainty and the corresponding anatomical regions. To this end, UR^2-MLLM is implemented through a three-stage pipeline. In Stage 1, we equip the model with uncertainty perception by fine-tuning it on a curated uncertainty-aware dataset, where hedging expressions (*e.g.*,"may", "appear", "could be") are aligned with their corresponding anatomical regions on the chest X-ray. This perception capability is a prerequisite for the model to later identify *when* a revisit is needed. In Stage 2, we construct a multimodal reasoning trajectory dataset of approximately 2,000 samples that explicitly supervises *when* and *where* to revisit, and introduce a detect-and-copy mechanism that inserts the indices of uncertainty-localized visual patches directly into the reasoning trace. Finally, in Stage 3, we refine this behavior with reinforcement learning using a visual grounding reward, encouraging revisited regions to remain aligned with the corresponding anatomical structures and reducing hallucinated content in the generated reports. Our contributions are as follows.

- We propose UR^2-MLLM, a novel framework to actively revisit visual evidence in an uncertainty-aware manner for RRG, explicitly modelling both when and where to revisit.
- We introduce a curated uncertainty-aware multimodal reasoning trajectory dataset, along with two core components, a patch-level visual injection mechanism and a visual grounding reward, integrated within a unified three-stage training pipeline.
- Extensive experiments on IU-Xray and MIMIC-CXR demonstrate that UR^2-MLLM achieves state-of-the-art performance, with consistent improvements in both efficiency, faithfulness, and interpretability.

2 Related Work

2.1 Uncertainty in RRG

Modeling diagnostic uncertainty is essential for radiology report generation, since clinical decisions depend on whether a finding is asserted with confidence or expressed with hedging. Existing approaches fall into three categories. **Probabilistic estimation** methods capture uncertainty via Monte Carlo dropout (Wang et al., 2024) or variational latent variable models (Najdenkoska et al., 2022). **Consistency-based** methods estimate uncertainty by measuring output variation across perturbed inputs (Zhang et al., 2024) or sampled generations (Wang et al., 2025). **Supervision-based** methods extract hedging cues from radiology reports and construct uncertainty-annotated data to train models that explicitly express diagnostic doubt (Wang et al., 2026; Rabaey et al., 2025). Our work falls into the third category, but goes beyond expressing uncertainty in text by using it as a trigger to actively revisit corresponding image regions during reasoning.

2.2 Reasoning in MLLMs for RRG

Recent advances in MLLMs have introduced explicit reasoning traces into radiology report genera-

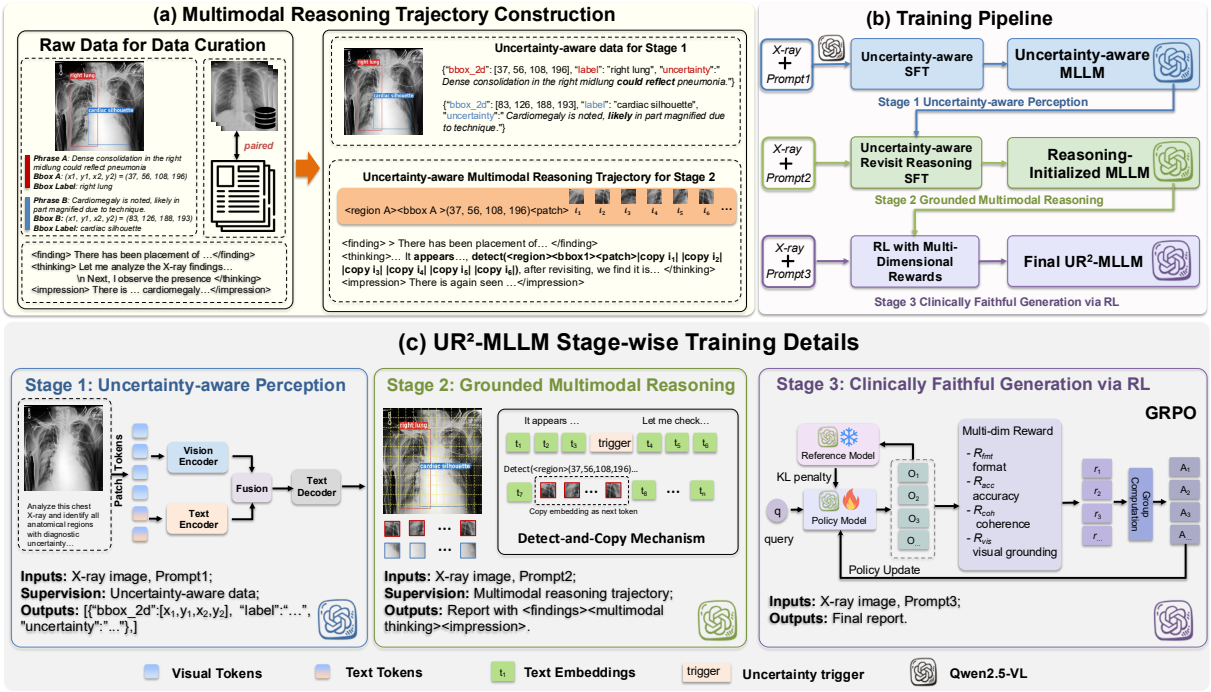


Figure 2: Overall framework of UR^2-MLLM, including (a) multimodal reasoning trajectory construction, (b) training pipeline overview, and (c) stage-wise training details.

tion. Existing approaches fall into two categories based on how visual evidence participates in reasoning. **Text-based reasoning** methods generate the reasoning trace purely in the textual space after the image has been encoded (Fan et al., 2025; Luo et al., 2025; Wang et al., 2026), which prevents the model from re-examining the image to verify or refine its claims. **Static visual reasoning** methods incorporate visual information, either as global image features or as region-level cues such as bounding boxes, alongside the reasoning process (Bannur et al., 2024; Jing et al., 2025); however, the visual context remains static once provided throughout generation, offering no mechanism to selectively revisit specific regions when needed. Inspired by how radiologists iteratively re-examine suspicious regions during diagnosis (Kundel et al., 1978), we design a framework in which visual revisits are dynamically triggered by diagnostic uncertainty, enabling the model to selectively attend to specific image regions throughout the reasoning trace.

3 Methodology

In this section, we present UR^2-MLLM, a framework for generating radiology reports that are visually grounded and faithful to clinical uncertainty. Emulating the radiologist's diagnostic workflow, our method consists of three training stages, namely *uncertainty-aware perception*, *grounded multimodal reasoning*, and *clinically faithful generation via reinforcement learning (RL)*, supported by a data curation pipeline that produces uncertainty-aware multimodal reasoning trajectories.

3.1 Task Formulation

Given a chest X-ray image $X \in \mathbb{R}^{H \times W \times 3}$ and a text instruction c (*i.e.*, a prompt asking the model to generate a structured radiology report), the goal of RRG is to produce a textual report $y = (y_1, y_2, \ldots, y_T)$ that accurately describes the visual findings and provides a clinical impression. Following standard practice, y typically contains a *Findings* section enumerating observed abnormalities and an *Impression* section summarizing diagnostic conclusions.

Conventional MLLM-based RRG methods learn a MLLM $\Omega_\theta(y \mid X, c)$, where θ denotes the model parameters, via maximum likelihood $\mathcal{L}_{\text{MLE}}(\theta) = -\sum_{t=1}^{T} \log \Omega_\theta(y_t \mid y_{1:t-1}, X, c)$. While effective for fluency, this objective treats y as a flat sequence of tokens and does not explicitly model the radiologist's iterative process of *visual evidence revisit* when uncertainty arises. To address this gap, UR^2-MLLM reformulates the objective as maximizing a multi-dimensional reward over uncertainty-aware

report generation:

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{(X,c)\sim\mathcal{D},\, y\sim\Omega_\theta(\cdot|X,c)} \big[\, R(y, X, c) \,\big], \quad (1)$$

where $\mathcal{D}$ denotes the training corpus of image–instruction pairs and $R(y, X, c) = \lambda_1 r_{\text{fmt}} + \lambda_2 r_{\text{acc}} + \lambda_3 r_{\text{coh}} + \lambda_4 r_{\text{vis}}$ is a multi-dimensional reward function as detailed in Section A.4.

3.2 Stage 1: Uncertainty-aware Perception

As shown in Fig. 2, Stage 1 equips the base MLLM with the ability to localize and articulate diagnostic uncertainty in chest X-ray images. We curate an uncertainty-aware dataset $\mathcal{D}_1$ built upon Chest ImaGenome (Wu et al., 2021), which provides bounding-box annotations for anatomical regions in MIMIC-CXR (Johnson et al., 2024). Following CURV (Wang et al., 2026), we associate each anatomical region with report phrases containing hedge markers (e.g., *may*, *likely*, *cannot exclude*), yielding a corpus of (image, bounding box, label, phrase) tuples. For each image, we aggregate its tuples into a structured target y and pair it with an instruction c that prompts the model to enumerate all regions expressing uncertainty. We then finetune the MLLM Ω_θ via supervised fine-tuning using the autoregressive negative log-likelihood, computed only over response tokens:

$$\mathcal{L}_{\text{S1}}(\theta) = -\,\mathbb{E}_{(X,c,y)\sim\mathcal{D}_1} \sum_{t=1}^{T} \log \Omega_\theta(y_t \mid y_{1:t-1},\, c,\, X). \quad (2)$$

The resulting model Ω_{θ_1} serves as the initialization for Stage 2, having acquired the perceptual prior that diagnostic hedging is grounded in specific anatomical regions. Details of the curation pipeline are provided in Section C.1 and Tabs. 8 and 15.

3.3 Stage 2: Grounded Multimodal Reasoning

Building on the uncertainty-aware perception acquired in Stage 1, Stage 2 trains the model to perform multimodal reasoning that revisits image regions whenever diagnostic uncertainty arises. We first describe how we construct grounded reasoning trajectories from the existing report corpus, and then introduce the detect-and-copy mechanism that enables the model to selectively attend to image patches during generation.

3.3.1 Multimodal Reasoning Trajectory Construction

We build our multimodal reasoning trajectories upon the text-based reasoning traces released by CURV (Wang et al., 2026), which provide sentence-level CoT annotations for each MIMIC-CXR report in the form of a ⟨findings⟩, ⟨thinking⟩, ⟨impression⟩ triplet.

To augment these text-only traces with visual revisit events, we scan the ⟨thinking⟩ block sentence by sentence and match each sentence against the phrase–bounding box pairs from Stage 1. When a sentence overlaps with an anatomical phrase, we **(i)** rewrite it to include a hedge cue and a region reference $\langle \text{obj}_n \rangle$ adjacent to the anatomical phrase, **(ii)** insert a structured `detect` call immediately after, and **(iii)** append a short revisit summary that resumes the reasoning conditioned on $\langle \text{obj}_n \rangle$.

To convert the bounding box b_n into a form consumable by the language model, we partition each input image into a regular $H' \times W'$ grid of non-overlapping patches and flatten them in row-major order, yielding a sequence of $P = H'W'$ patches indexed from 1 to P. The bounding box b_n is then mapped to the set of patch indices $\mathcal{P}_n \subseteq \{1, \ldots, P\}$ that fall within b_n. These indices are inserted into the `detect` call block alongside the anatomical label, enabling the model to reference the revisited region at patch-level granularity. The resulting trajectory is represented as follows:

$$\underbrace{s_{1:i-1}}_{\text{plain}} \rightarrow \underbrace{\tilde{s}_i}_{\text{hedged}} \rightarrow \underbrace{\texttt{detect}(g_n, b_n, \mathcal{P}_n)}_{\text{revisit}} \rightarrow \underbrace{r_n}_{\text{summary}} \rightarrow, \quad (3)$$

where g_n and b_n denote the anatomical label and bounding box of the revisited region, $\tilde{s}_i$ is the rewritten hedged sentence, and r_n is the post-revisit summary, both anchored to the region tag $\langle \text{obj}_n \rangle$. Sentences matching no anatomical phrases remain unchanged. This formulation establishes a behavioral pattern in which uncertainty triggers grounded visual revisits, and the model learns to continue reasoning after referencing a specific image region. The dataset $\mathcal{D}_2 = \{(X, c, y)\}$ follows the same instruction–response format as Stage 1, with y being the full ⟨findings⟩–⟨thinking⟩–⟨impression⟩ sequence containing region tags, bounding boxes, and patch indices.

3.3.2 Detect-and-Copy Mechanism

We introduce a *detect-and-copy* mechanism, inspired by pointer-generator networks (See et al., 2017; Vinyals et al., 2015), which extends the language head to emit image patch indices directly within the reasoning trace as shown in Fig. 3.

Let $\mathbf{H}_t \in \mathbb{R}^d$ denote the LLM hidden state at decoding step t, and $\mathbf{V} \in \mathbb{R}^{P\times d}$ the visual em-

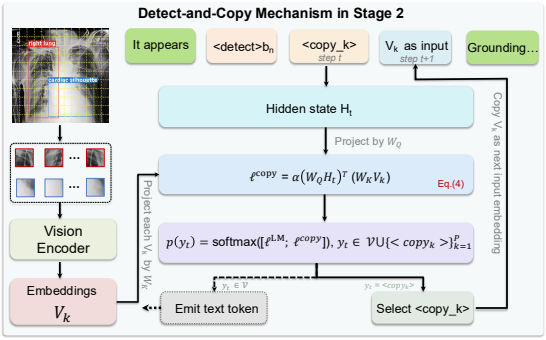


Figure 3: Details for Detect-and-Copy mechanism. $\mathcal{V}$ denotes the vocabulary and P denotes #patches.

beddings of the P image patches produced by the vision encoder. We extend the model vocabulary with P copy tokens $\{\langle \text{copy}_k \rangle\}_{k=1}^{P}$, each acting as a learned pointer to one image patch. Linear projections W_Q and W_K map hidden states and visual embeddings into a shared space, and the logit for copy token k at step t is the scaled inner product:

$$\ell_{t,k}^{\text{copy}} = \alpha \cdot \left(W_Q \mathbf{H}_t\right)^{\top} \left(W_K \mathbf{V}_k\right), \tag{4}$$

where α is a learnable temperature. These copy logits are concatenated with the standard language-model logits and trained jointly under a single cross-entropy objective, allowing the model to decide at each step whether to emit a text token or a patch pointer. Unlike pointer-generator networks (See et al., 2017), no copy gate is required, since the language-model logits and copy logits are defined over disjoint output spaces and normalized by a single softmax. Within a `detect` call, the model emits the bounding box b_n followed by the copy tokens corresponding to the patches in $\mathcal{P}_n$, producing a learned hard-attention mask anchored to the revisited region. Since the copy logits share their key space with the visual embeddings, subsequent reasoning tokens are naturally grounded in the visual content of the revisited patches. Starting from Ω_{θ_1}, we train Stage 2 with the response-masked autoregressive negative log-likelihood over the extended vocabulary:

$$\mathcal{L}_{\text{S2}}(\theta) = -\,\mathbb{E}_{(X,c,y)\sim\mathcal{D}_2} \sum_{t=1}^{T} \log \Omega_\theta(y_t \mid y_{1:t-1},\, c,\, X). \tag{5}$$

The resulting model Ω_{θ_2} serves as the reasoning-initialized MLLM for Stage 3.

3.4 Stage 3: Clinically Faithful Generation via RL

In this stage, the training data consists of image–report pairs $(X, y^*) \in \mathcal{D}_3$ drawn from MIMIC-CXR, where y^* serves as the ground-truth reference for reward computation. While Stage 2 teaches the model to invoke `detect` calls, it provides no signal on whether the revisited region is semantically consistent with the reasoning that follows. To close this loop, we introduce a *visual grounding reward* r_{vis} that directly supervises the quality of visual revisits. From each trajectory τ, we extract the set of revisit events $\mathcal{E}(\tau) = \{(b_e, \text{txt}_e)\}_e$, where b_e is the bounding box emitted within a `detect` call and txt_e is the revisit summary sentence emitted immediately after. Two frozen encoders from BiomedCLIP (Zhang et al., 2023), ψ_v and ψ_t, embed the image patch I_{b_e} and the text txt_e into a shared space. Since cosine similarity between a region and text can be inflated by domain-level visual–textual correlation, we introduce a random-region baseline to calibrate the score. Specifically, we sample K random bounding boxes $\{b_e^k\}_{k=1}^{K}$ matching b_e in area and average their similarity scores. The visual-grounding reward r_{vis} is then defined as:

$$r_{\text{vis}}(\tau) = \frac{1}{|\mathcal{E}(\tau)|} \sum_{e \in \mathcal{E}(\tau)} \Big[s(b_e, \text{txt}_e) - \bar{s}(b_e, \text{txt}_e) \Big], \tag{6}$$

where $s(b, \text{txt}) = \cos\big(\psi_v(I_b), \psi_t(\text{txt})\big)$ is the cosine similarity between the visual encoding of the image patch I_b and the text encoding of txt, and $\bar{s}(b, \text{txt}) = \frac{1}{K}\sum_{k=1}^{K} s(b^k, \text{txt})$ is the average similarity over K random bounding boxes $\{b^k\}_{k=1}^{K}$ of the same area as b. We set $r_{\text{vis}}(\tau) = 0$ if $\mathcal{E}(\tau) = \emptyset$. The subtraction centers the signal around zero, so that the reward measures the *specificity* of the predicted bounding box rather than generic domain-level similarity.

We optimize the model with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), taking Ω_{θ_2} as both the initial policy and the frozen reference Ω_{ref}. The visual grounding reward is combined with format adherence r_{fmt}, clinical accuracy r_{acc}, and reasoning coherence r_{coh} into a composite reward $R(\tau) = \sum_k w_k\, r_k(\tau)$, which drives the objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\,\mathbb{E}\Big[\tfrac{1}{G}\textstyle\sum_{i=1}^{G} \min(\rho_i A_i,\, \text{clip}_\epsilon(\rho_i) A_i)\Big] + \beta\, \text{KL}[\Omega_\theta \,\|\, \Omega_{\text{ref}}], \tag{7}$$

where A_i is the group-relative advantage of trajectory τ_i and $\rho_i = \Omega_\theta(\tau_i \mid x)/\Omega_{\theta_{\text{old}}}(\tau_i \mid x)$. Here, $\text{clip}_\epsilon(\rho_i) = \min(\max(\rho_i, 1-\epsilon), 1+\epsilon)$ denotes clipping ρ_i to $[1-\epsilon, 1+\epsilon]$. The resulting model, denoted UR^2-MLLM, generates radiology reports that are visually grounded and clinically faithful.

4 Experiments

4.1 Experimental Setup

Dataset. We curate training data from MIMIC-CXR (Johnson et al., 2024) and Chest ImaGenome (Wu et al., 2021), building an uncertainty-aware perception corpus $\mathcal{D}_1$ for Stage 1 and a multimodal reasoning trajectory corpus $\mathcal{D}_2$ for Stage 2, while Stage 3 reuses MIMIC-CXR dataset without additional annotation. We further test out-of-distribution generalization on the unseen IU X-ray (Demner-Fushman et al., 2016). More details are provided in Section A.1.

Evaluation Metrics. We evaluate the generated reports from natural language generation (NLG) and clinical efficacy (CE) metrics. CE is assessed by CheXbert (Smit et al., 2020) (Accuracy, Macro-F1, Micro-F1) and RadGraph (Jain et al., 2021) (Entity F1, complete F1). In Tab. 2 and Tab. 4, we denote Ma-F1, Mi-F1 and Ent.F1 as Macro F1, Micro F1 and Entity F1, respectively. More descriptions are in Section A.2.

Implementation Details. We initialize all stages from Qwen2.5-VL-3B (Team, 2025b) and train UR^2-MLLM on 4 NVIDIA A6000 GPUs. Stages 1 and 2 use supervised fine-tuning, while Stage 3 applies GRPO. Full training and inference hyperparameters are provided in Section A.3.

4.2 Comparison Results

We evaluate UR^2-MLLM on the MIMIC-CXR (Johnson et al., 2024) test split against a diverse set of baselines, including general-purpose MLLMs (LLaVA-1.5-7B (Liu et al., 2024a), Qwen2.5-VL-3B (Team, 2025b), Gemini 2.5 Pro (Team, 2025a)), medical-domain MLLMs (HuatuoGPT-Vision-7B (Chen et al., 2024), LLaVA-1.5-7B-SFT-CXR), the grounded RRG model MAIRA-2 (Bannur et al., 2024) and CURV (Wang et al., 2026) as the closest reasoning-oriented baseline. As shown in Tab. 1, UR^2-MLLM achieves the best performance on five out of six NLG metrics, with BLEU-1, BLEU-2, BLEU-3, BLEU-4, and ROUGE-L all surpassing the strongest baseline CURV. On METEOR, our model scores 28.66, slightly below CURV's 30.43 but remaining competitive against all other MLLM baselines. On the clinical efficacy metrics in Tab. 2, UR^2-MLLM achieves 78.67 CheXbert Acc., 13.96 Ma-F1, 62.96 Mi-F1, 27.72 RadGraph Ent.F1, and 21.47 RadGraph F1, outperforming CURV across all 5 metrics. The improvement is particularly notable on the clinically-oriented metrics such as RadGraph F1, suggesting that grounding the reasoning trace in visual evidence through adaptive revisit produces reports that are more clinically faithful, not merely more fluent.

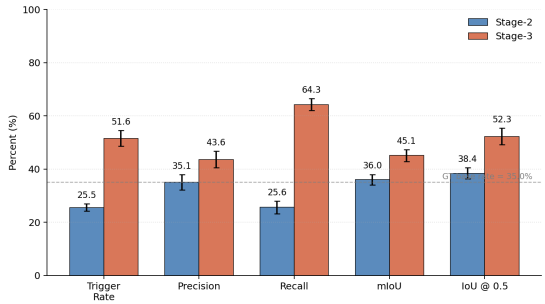


Figure 4: Revisit-quality on the MIMIC-CXR.

4.3 Out-of-Distribution Evaluation

To assess generalization beyond the training distribution, we evaluate UR^2-MLLM on IU X-ray (Demner-Fushman et al., 2016), which is unseen during all three training stages and differs from MIMIC-CXR in patient population, acquisition protocol, and reporting style. As shown in Tab. 3, UR^2-MLLM achieves the best performance across all six NLG metrics, reaching 31.81 BLEU-1, 7.11 BLEU-4, 40.75 METEOR, and 41.20 ROUGE-L, consistently outperforming CURV and all other MLLM baselines. On the clinical efficacy side (Tab. 4), UR^2-MLLM attains 92.58 CheXbert Acc. and 8.45 Ma-F1, both the highest among all baselines, along with the top RadGraph Ent.F1 (37.87) and RadGraph F1 (25.95). The only metric on which our model does not take the lead is CheXbert Mi-F1 (74.15 vs. CURV's 74.36), with the gap being marginal. These results demonstrate that the gains brought by adaptive visual revisit are not specific to the MIMIC-CXR distribution, but transfer to an unseen dataset, indicating that the model has learned a transferable strategy of grounding reasoning in visual evidence rather than overfitting to dataset-specific patterns.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
LLaVA-1.5-7B (Liu et al., 2024a)	19.09	7.46	2.81	1.25	19.16	18.36
LLaVA-1.5-7B-SFT-CXR	22.58	15.06	9.43	6.13	25.71	28.09
HuatuoGPT-Vision-7B (Chen et al., 2024)	19.33	9.42	4.64	1.93	26.01	20.78
MAIRA-2 (Bannur et al., 2024)	24.94	14.12	9.01	6.14	26.78	28.65
Qwen2.5-VL-3B (Team, 2025b)	13.09	5.42	2.08	0.89	20.81	15.23
Gemini 2.5 Pro (Team, 2025a)	12.54	5.20	2.25	1.05	21.19	15.01
CURV (Wang et al., 2026)	25.38	15.58	9.85	6.18	**30.43**	31.19
UR^2-MLLM	**27.79**	**16.20**	**10.11**	**6.73**	28.66	**32.67**

Table 1: NLG metrics for radiology report generation across different models on MIMIC-CXR(**Best**, Second-best).

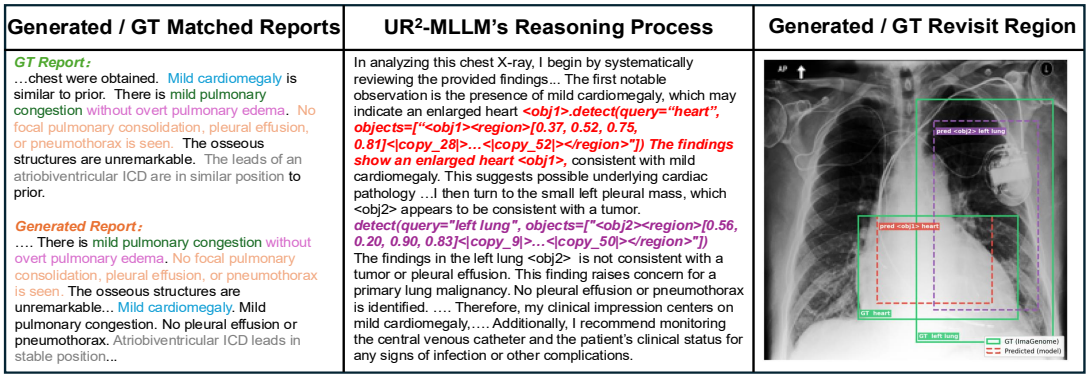


Figure 5: A success case of UR^2-MLLM. Left: ground-truth and generated reports with matched findings highlighted. Middle: reasoning trace where uncertainty triggers `detect` calls. Right: predicted bounding boxes (red, purple) versus ground-truth (green) from Chest ImaGenome.

Model	CheXbert			RadGraph	
	Acc.	Ma-F1	Mi-F1	Ent.F1	F1
LLaVA-1.5-7B	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	71.15	5.34	48.62	15.92	9.06
MAIRA-2	67.39	6.34	46.53	25.01	17.05
Qwen2.5-VL-3B	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	76.93	5.22	57.12	25.95	19.54
UR^2-MLLM	**78.67**	**13.96**	**62.96**	**27.72**	**21.47**

Table 2: Clinical accuracy metrics on the MIMIC-CXR dataset (**Best**, Second-best).

4.4 Ablation Study

To validate each design choice in UR^2-MLLM, we conduct an ablation study on MIMIC-CXR, examining the contribution of each training stage, dynamic visual revisit versus static visual injection, and the effect of the visual grounding reward r_{vis}, as reported in Tab. 5. Starting from the Qwen2.5-VL backbone (Base), Stage 1 alone (+S1) slightly degrades NLG metrics, since this stage focuses on structured uncertainty enumeration rather than full report generation, but substantially improves clinical F1 (4.75 $\rightarrow$ 12.35). Adding Stage 2 (+S1+S2) yields a large NLG gain as the model produces complete reasoning trajectories, and Stage 3 further pushes performance to 27.79 BLEU-1, 28.66 METEOR, and 13.96 clinical F1. To isolate the value of dynamic revisit, we further consider two Stage 2 variants: w/o visual revisit, which disables the detect-and-copy mechanism, and w/ static vision, where the full image together with the target bounding box is provided as a fixed visual context that the reasoning attends to via cross-attention throughout, without dynamically revisiting specific regions. The w/o visual revisit variant drops sharply on both NLG (BLEU-1: 10.72) and clinical F1 (3.87), and the w/ static vision variant reaches only 16.13 BLEU-1 and 6.13 clinical F1, still far behind our dynamic revisit design. Finally, removing r_{vis} from Stage 3 degrades both NLG and clinical efficacy (BLEU-1: 2.53$\downarrow$, clinical F1: 0.96 $\downarrow$, Recall: 1.75$\downarrow$), confirming that supervising *where* the model looks is not sufficient on its own and that the reward plays a complementary role in aligning the revisited region with the textual claim. Overall, these results validate that each stage and design component is essential, and that dynamic visual revisit anchored by uncertainty and reinforced by a visual grounding reward is the key ingredient

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
LLaVA-1.5-7B	16.52	6.60	3.00	1.40	19.65	17.48
LLaVA-1.5-7B-SFT-CXR	21.42	12.95	8.03	5.20	23.24	26.40
HuatuoGPT-Vision-7B	19.33	10.70	6.28	2.81	31.02	23.42
MAIRA-2	26.37	15.60	9.64	6.03	25.52	31.18
Qwen2.5-VL-3B	11.38	5.05	2.28	1.05	21.65	15.01
CURV	29.23	18.76	12.08	6.86	38.30	39.08
UR^2-MLLM	**31.81**	**19.49**	**12.94**	**7.11**	**40.75**	**41.20**

Table 3: NLG metrics for radiology report generation on the IU X-ray dataset (**Best**, Second-best).

Model	CheXbert			RadGraph	
	Acc.	Ma-F1	Mi-F1	Ent.F1	F1
LLaVA-1.5-7B	72.03	4.66	46.81	13.37	8.76
LLaVA-1.5-SFT	76.34	5.84	53.33	16.19	10.31
HuatuoGPT-V-7B	89.89	5.57	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	91.56	5.86	**74.36**	36.99	25.65
UR^2-MLLM	**92.58**	**8.45**	74.15	**37.87**	**25.95**

Table 4: Clinical accuracy metrics on the IU X-ray dataset (**Best**, Second-best).

behind UR^2-MLLM's performance.

4.5 Visual Reasoning Evaluation

To assess whether UR^2-MLLM genuinely learns when and where to revisit visual evidence rather than imitating the surface form of reasoning traces, we evaluate its revisit behavior on the MIMIC-CXR test split (n=3,858). We report the trigger rate, the precision and recall of the triggering decision, and the spatial accuracy of the revisited region measured by mIoU and IoU@0.5. The ground-truth base rate of samples genuinely requiring a revisit is 35.0%. As shown in Fig. 4, Stage 3 consistently improves over Stage 2 across all five metrics. The trigger rate rises from 25.5% to 51.6%, crossing the base rate and shifting the model from under-triggering to actively invoking revisits. The recall improves substantially from 25.6% to 64.3%. These indicate that the reinforcement learning stage instills a visually grounded revisit policy rather than a purely linguistic one. While these improvements are clear, the absolute precision and spatial overlap remain limited.

4.6 Visual Revisit Validation

Localization overlap alone does not establish that the statement after a revisit describes the selected region. We therefore score the semantic coherence between each generated revisit summary and its corresponding image crop on a five-point scale, and compare the selected region with a same-size random-region control. As shown in Tab. 6, coherence rises from 2.78 after Stage 2 SFT to 3.81 for the full model, whereas the random-region control scores only 1.23. This separation indicates that the post-revisit text is specific to the selected visual evidence rather than a generic continuation. The judging protocol is provided in Section E.3.

We further test whether the revisited embeddings causally affect subsequent generation. For 200 test samples containing a revisit, we continue decoding under four conditions: clean selected embeddings (Clean revisit), the same number of random non-revisited patches (Random patches), zeroed selected embeddings (Zeroed revisit), and embeddings from an unrelated anatomical region (Mismatched region). Tab. 7 reports CheXbert Macro-F1 and the fraction of post-revisit summaries that change relative to clean decoding. Zeroing the selected embeddings produces the largest F1 drop (2.51) and changes 82.0% of summaries; mismatched regions reduce F1 by 1.31 and change 73.0%. In contrast, perturbing the same number of random non-revisited patches causes only a 0.42 F1 drop. These controlled interventions show that generation depends preferentially on the selected revisit embeddings rather than on arbitrary visual-token corruption. Full intervention details are given in Section E.4.

4.7 Case Study

Fig. 5 presents a representative success case from MIMIC-CXR. The report generated by UR^2-MLLM closely aligns with the ground truth across all key findings, including mild cardiomegaly, mild pulmonary congestion, the absence of pleural effusion or pneumothorax, and the atriobiventricular ICD. The middle column further reveals the model's reasoning process: when diagnostic uncertainty arises, the model triggers `detect` calls to

Model	NLG						CE		
	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	F1	Rec.	Prec.
Base (Qwen2.5-VL-3B)	13.09	5.42	2.08	0.89	20.81	15.23	4.75	-	-
+ S1	9.88	4.31	1.91	1.02	20.33	10.74	12.35	28.06	13.47
+ S2	23.66	13.88	8.91	5.96	21.77	18.69	11.66	24.50	14.68
+ S1 + S2 (w/o visual revisit)	10.72	5.22	2.43	1.10	18.57	14.59	3.87	-	-
+ S1 + S2 (w/ static vision)	16.13	5.01	5.19	1.98	17.10	14.97	6.13	21.30	14.96
+ S1 + S2	26.98	15.61	9.75	6.44	23.36	21.36	13.55	32.34	17.25
+ S1 + S2 + S3 (w/o r_{vis})	25.26	13.94	8.03	5.07	24.96	16.11	13.00	30.64	**26.09**
+ S1 + S2 + S3 (**UR^2-MLLM**)	**27.79**	**16.20**	**10.11**	**6.73**	**28.66**	**32.67**	**13.96**	**32.39**	18.03

Table 5: Ablation study of UR^2 on MIMIC-CXR (**Best**, Second-best). S1, S2, S3 denotes Stage1, Stage2, and Stage3, respectively. F1 here denotes CheXbert Macro F1.

Condition	Coherence (1–5)
Random-region control	1.23
Stage 2 SFT	2.78
Full model	**3.81**

Table 6: Region-text coherence on MIMIC-CXR revisit events. Higher is better. Three conditions are defined as: (i) a same-size random crop paired with the unchanged summary, (ii) the crop selected by the Stage 2 SFT model, and (iii) the crop selected by the full Stage 3 model.

Condition	CheXbert F1	ΔF1	Changed (%)
Clean revisit	14.59	0.00	–
Random patches	14.17	0.42	43.0
Zeroed revisit	12.08	2.51	82.0
Mismatched region	13.28	1.31	73.0

Table 7: Causal patch-perturbation study on 200 MIMIC-CXR test samples. ΔF1 is the absolute drop from clean decoding; "Changed" is the percentage of regenerated post-revisit summaries that differ from the clean condition.

revisit specific anatomical regions, first the heart ($\langle \text{obj}_1 \rangle$) and then the left lung ($\langle \text{obj}_2 \rangle$). As shown in the right column, the predicted bounding boxes (red and purple) are spatially well-aligned with the Chest ImaGenome ground-truth boxes (green), confirming that the model attends to the correct anatomical structures during revisit. Additional cases are provided in Section F.

5 Conclusion

We present UR^2-MLLM, an uncertainty-aware multimodal large language model for radiology report generation that emulates how radiologists iteratively revisit suspicious regions during diagnosis. Unlike prior methods that rely on static visual evidence, UR^2-MLLM treats diagnostic uncertainty as an explicit trigger for visual revisit, deciding *when* to look again and *where* to focus as reasoning unfolds. The model progressively learns to perceive uncertainty, ground reasoning in localized visual evidence, and refine revisit behavior via reinforcement learning. Across in-distribution and out-of-distribution benchmarks, UR^2-MLLM achieves consistent improvements on both linguistic and clinical efficacy metrics, with further analysis showing that these gains stem from genuinely grounded visual revisit rather than imitation of surface reasoning patterns. We hope this work encourages further research on visually grounded, uncertainty-aware reasoning for trustworthy medical report generation.

Limitations

While UR^2-MLLM demonstrates the value of uncertainty-triggered visual revisit, several limitations remain. First, our notion of diagnostic uncertainty is derived from textual hedging cues, which provides only a coarse and weakly constrained signal. A more clinically grounded formulation of uncertainty, potentially informed by radiologist confidence ratings or calibrated probability estimates, could yield a stronger and more reliable trigger. Second, the multimodal reasoning trajectory corpus used in Stage 2 is relatively small (∼2K trajectories) and is partly derived from LLM-rewritten reasoning traces, which may introduce upstream biases. Scaling up the dataset with higher-quality and expert-validated trajectories is an important direction for future work. Third, the spatial grounding in our pipeline relies on the silver-standard anatomical annotations of Chest ImaGenome, which constrains the precision of the revisit supervision and limits applicability to settings where such region-level annotations are unavailable. Finally, our evaluation is confined to chest X-ray benchmarks with automatic metrics. The absence of external clinical

validation and human-in-the-loop assessment by radiologists leaves the real-world clinical reliability of the generated reports an open question. We view UR^2-MLLM as a meaningful but preliminary step toward uncertainty-aware visual reasoning, and addressing these limitations is essential for clinical deployment.

Ethics Considerations

This research complies with established ethical standards. All datasets employed are publicly available and utilized solely for their intended research purposes. These datasets contain no personally identifiable information or sensitive content, thereby posing no risks to privacy or confidentiality. In addition, the study did not involve human subjects or annotators.

References

Satanjeev Banerjee and Alon Lavie. 2005. 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.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. 2024. MAIRA-2: grounded radiology report generation. *CoRR*, abs/2406.04449.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. 2024. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370, Miami, Florida, USA. Association for Computational Linguistics.

Yucheng Chen, Yang Yu, Yufei Shi, Conghao Xiong, Xulei Yang, and Si Yong Yeo. 2026. Riha: Report-image hierarchical alignment for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*.

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

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

Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. *arXiv preprint arXiv:2504.20930*.

Amaan Izhar, Norisma Idris, and Nurul Japar. 2025. Medical radiology report generation: A systematic review of current deep learning methods, trends, and future directions. *Artif. Intell. Medicine*, 168:103220.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Matthew Lungren, Andrew Ng, Curtis Langlotz, and Pranav Rajpurkar. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *PhysioNet*. Version 1.0.0.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 2607–2615.

Peiyuan Jing, Kinhei Lee, Zhenxuan Zhang, Huichi Zhou, Zhengqing Yuan, Zhifan Gao, Lei Zhu, Giorgos Papanastasiou, Yingying Fang, and Guang Yang. 2025. Reason like a radiologist: Chain-of-thought and reinforcement learning for verifiable report generation. *Medical Image Analysis*, page 103910.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. 2024. MIMIC-CXR Database. *PhysioNet*. Version 2.1.0.

Harold L Kundel, Calvin F Nodine, and Dennis Carmody. 1978. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative radiology*, 13(3):175–181.

Yilin Li, Chao Kong, Guosheng Zhao, and Zijian Zhao. 2025. Automatic radiology report generation with deep learning: a comprehensive review of methods and advances. *Artificial Intelligence Review*, 58(11):344.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024b. Medcot: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389.

Yihong Luo, Wenwu He, Zhuo-Xu Cui, and Dong Liang. 2025. Teaching ai stepwise diagnostic reasoning with report-guided chain-of-thought learning. *arXiv preprint arXiv:2509.06409*.

Ivona Najdenkoska, Xiantong Zhen, Marcel Worring, and Ling Shao. 2022. Uncertainty-aware report generation for chest x-rays by variational topic inference. *Medical Image Analysis*, 82:102603.

Chee Ng, Liliang Sun, and Shaoqing Tang. 2025. X-ray-cot: Interpretable chest x-ray diagnosis with vision-language models via chain-of-thought reasoning. *arXiv preprint arXiv:2508.12455*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Paloma Rabaey, Jong Hak Moon, Jung-Oh Lee, Min Gwan Kim, Hangyul Yoon, Thomas Demeester, and Edward Choi. 2025. Modeling clinical uncertainty in radiology reports: from explicit uncertainty markers to implicit reasoning pathways. *arXiv preprint arXiv:2511.04506*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. 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.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, Linjie Li, Yu Cheng, Heng Ji, Junxian He, and Yi R. Fung. 2025. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *Preprint*, arXiv:2506.23918.

Gemini Team. 2025a. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Qwen Team. 2025b. Qwen3-vl technical report. *CoRR*, abs/2511.21631.

A Van der Gijp, CJ Ravesloot, H Jarodzka, MF Van der Schaaf, IC Van der Schaaf, Jan PJ van Schaik, and Th J Ten Cate. 2017. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. *Advances in Health Sciences Education*, 22(3):765–787.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

Chenyu Wang, Weichao Zhou, Shantanu Ghosh, Kayhan Batmanghelich, and Wenchao Li. 2025. Semantic consistency-based uncertainty quantification for factuality in radiology report generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1739–1754.

Yixin Wang, Zihao Lin, Zhe Xu, Haoyu Dong, Jie Luo, Jiang Tian, Zhongchao Shi, Lifu Huang, Yang Zhang, Jianping Fan, and Zhiqiang He. 2024. Trust it or not: Confidence-guided automatic radiology report generation. *Neurocomputing*, 578:127374.

Ziao Wang, Sixing Yan, Kejing Yin, Xiaofeng Zhang, and William K Cheung. 2026. Curv: Coherent uncertainty-aware reasoning in vision-language models for x-ray report generation. *Advances in Neural Information Processing Systems*, 38:93686–93713.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, Tanveer Syeda-Mahmood, and Mehdi Moradi. 2021. Chest ImaGenome Dataset. *PhysioNet*. Version 1.0.0.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. 2024. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Kaur Bagga, Robert Tinn, Sam Preston, Rajesh N. Rao, Mu-Hsin Wei, Naveen Valluri, Cliff Wong, Matthew P. Lungren, Tristan Naumann, and Hoifung Poon. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *Preprint*, arXiv:2303.00915.

Appendix

A Experimental Setup

A.1 Dataset

We curate our training data from two publicly available resources. MIMIC-CXR (Johnson et al., 2024) is a large-scale chest X-ray dataset containing over 370,000 images paired with free-text radiology reports. Chest ImaGenome (Wu et al., 2021) provides anatomical bounding box annotations and structured scene graphs for MIMIC-CXR studies. Built upon these two resources, we construct two stage-specific corpora: an uncertainty-aware perception corpus $\mathcal{D}_1$ containing approximately 101k samples used in Stage 1, and a multimodal reasoning trajectory corpus $\mathcal{D}_2$ containing around 2,000 curated trajectories used in Stage 2. Stage 3 reinforcement learning samples prompts from MIMIC-CXR without requiring additional annotation. All experiments adopt the train/test split defined by PromptMRG (Jin et al., 2024), with 3,858 studies held out for evaluation. To further assess out-of-distribution generalization, we additionally evaluate on IU X-ray (Demner-Fushman et al., 2016), a smaller chest X-ray dataset that the model has not been exposed to during training. The statistic details for our curation datasets can be found in Tab. 8.

A.2 Evaluation Metrics

We evaluate the generated reports along two complementary axes. Natural language generation (NLG) quality is measured against the ground-truth report using BLEU-1/2/3/4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE-L (Lin, 2004). Clinical efficacy (CE) evaluates the factual agreement on the 14 CheXpert pathologies, with predicted labels extracted by the CheXbert labeler (Smit et al., 2020). We report Accuracy, Macro-F1, and Micro-F1. We additionally report Entity F1 and complete F1 from RadGraph (Jain et al., 2021) to assess clinical entity alignment.

A.3 Implementation Details

We initialize all stages from Qwen2.5-VL-3B (Team, 2025b) and train UR^2-MLLM on 4 NVIDIA A6000 GPUs with BF16 mixed precision. We use AdamW with a cosine learning rate schedule and 5% warm-up. Stages 1 and 2 are trained via supervised fine-tuning for 3 epochs each, with a learning rate of 1×10^{-5} and an effective batch size of 32. Stage 3 further refines the model via GRPO for 1 epoch over a sampled subset of MIMIC-CXR prompts, with $G = 4$ trajectories per prompt, a KL coefficient $\beta = 0.04$, and clip range $\epsilon = 0.2$. At inference, we use nucleus sampling with temperature $T = 0.7$ and top-$p = 0.9$. Our code will be released upon acceptance.

A.4 Reward Definitions

Tab. 9 consolidates the four Stage 3 reward components that are introduced in Section 3.4. The format component verifies the required report schema and section order. The clinical-accuracy component compares CheXbert labels extracted from the generated and reference reports over the 14 CheXpert observations. The coherence component evaluates whether the reasoning is internally consistent and supports the generated Findings and Impression. The visual component is the calibrated region–text similarity in Eq. (6); subtracting a same-area random-region baseline makes it reward region-specific evidence rather than generic chest-X-ray/text compatibility. All components are computed per sampled trajectory before group-relative normalization in GRPO.

A.5 Worked Detect-and-Copy Decoding Example

The following schematic example makes the inference-time data flow explicit. Suppose the model generates a hedged statement about the left lower lung and then opens a revisit call:

```
. . . an opacity may involve <obj1> the
left lower lung.
detect(<obj1>,
<region>[0.08,0.46,0.48,0.91]</region>,
<copy31><copy32><copy47><copy48>)
```

The bounding box is mapped to four grid cells, so the decoder emits the corresponding copy tokens under the joint softmax in Eq. (4). Each emitted copy token is then replaced by its projected visual embedding $(\mathbf{V}_{31}, \mathbf{V}_{32}, \mathbf{V}_{47}, \mathbf{V}_{48})$ before textual decoding resumes. The next hidden state therefore conditions on these local embeddings and may continue, for example, with a cautious post-revisit summary about the opacity. The copied vectors provide local evidence; they do not force a predetermined textual phrase. This distinction is also why the perturbations in Tab. 7 can change the downstream summary.

Dataset	Split	Image-level	Item-level
Stage 1 SFT data			
uncertainty-aware dataset	full	101,457 images	542,834 bbox items
(∼5.35 bbox / image, median 4, max 39)	—	—	—
Stage 2 SFT data			
multimodal_reasoning_trajectory	full	1,909 images	3,172 revisit anchors
Type A (with detect() calls)	—	1,292 images	3,172 anchors
Type B (no detect() calls)	—	617 images	0 anchors
train / val (5% holdout)	train / val	1,814 / 95	—
Stage 3 RL data			
MIMIC-CXR ((Johnson et al., 2024))	train	227,835 samples	—
	test	3,858 samples	—
Out-of-domain evaluation			
IU-Xray ((Demner-Fushman et al., 2016))	train	2,069 samples	—
	val	296 samples	—
	test	590 samples	—

Table 8: Dataset sizes used by UR^2-MLLM. "Image-level" counts the unique image-report pairs that become training (or evaluation) samples; "item-level" counts the finer-grained annotations derived from each pair, since a single MIMIC-CXR study yields multiple bbox-uncertainty annotations during Stage 1 curation and a single ImaGenome-anchored sentence yields multiple revisit regions during Stage 2 curation, so item-level counts are strictly larger than image-level counts wherever applicable. Stage 3 uses MIMIC-CXR image-report pairs without additional annotation. IU-Xray is used only for out-of-domain evaluation; its "item-level" column counts the individual frontal and lateral DICOM images, since a single report covers both views.

Term	Weight	Input	Computation
r_{fmt}	λ_1	generated report	Validity of the `<findings>` → `<thinking>` → `<impression>` schema and section order.
r_{acc}	λ_2	generated/reference reports	CheXbert multi-label agreement over the 14 CheXpert observations.
r_{coh}	λ_3	reasoning and final report	Consistency of `<thinking>` with the generated Findings and Impression.
r_{vis}	λ_4	selected crop and revisit summary	BiomedCLIP cosine similarity minus the mean similarity of K same-area random crops (Eq. (6)).

Table 9: Stage 3 reward definitions. The composite reward is $R = \lambda_1 r_{\text{fmt}} + \lambda_2 r_{\text{acc}} + \lambda_3 r_{\text{coh}} + \lambda_4 r_{\text{vis}}$.

B Backbone Scaling

To examine whether UR^2-MLLM scales to a larger backbone, we replace Qwen2.5-VL-3B-Instruct with Qwen2.5-VL-7B-Instruct and train the complete Stage 1–3 pipeline. Our Stage 1 and Stage 2 corpora are curated once using a frozen Qwen2.5-7B-Instruct model, which is employed offline for report parsing, uncertainty extraction, and reasoning-trajectory rewriting. The curation model processes report-derived text and associated anatomical annotations but does not receive the radiographs, and is therefore functionally distinct from the multimodal backbone used for report generation. To ensure a controlled comparison, the 3B and 7B variants use exactly the same frozen curated datasets, stage-wise supervision, data splits, and evaluation protocol; no additional supervision is regenerated for the 7B experiment.

As shown in Tab. 10, increasing the multimodal backbone from 3B to 7B improves all reported NLG and clinical efficacy metrics. In particular, BLEU-1, BLEU-2, and BLEU-3 increase from 27.79, 16.20, and 10.11 to 29.81, 19.08, and 12.11, respectively, while METEOR and ROUGE-L improve from 28.66 and 32.67 to 29.12 and 33.09. The larger backbone also raises CheXbert Macro-F1 from 13.96 to 15.34, recall from 32.39 to 33.18, and precision from 18.03 to 20.07. These consistent gains indicate that the proposed uncertainty-aware visual-revisit pipeline benefits from increased backbone capacity. Although the curation model and both multimodal backbones belong to the Qwen2.5 family, they serve distinct offline-curation and image-conditioned generation roles; accordingly, this experiment demonstrates backbone scaling within the same model family rather than cross-family generalization.a larger backbone.

Backbone	NLG						CE		
	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	F1	Rec.	Prec.
Qwen2.5-VL-3B	27.79	16.20	10.11	6.73	28.66	32.67	13.96	32.39	18.03
Qwen2.5-VL-7B	29.81	19.08	12.11	7.09	29.12	33.09	15.34	33.18	20.07

Table 10: Backbone-scale comparison of the full UR^2-MLLM pipeline on the MIMIC-CXR test set. Both variants use the complete Stage 1–3 pipeline and the same evaluation protocol. F1 denotes CheXbert Macro-F1.

C Prompt details

C.1 Data Curation Prompts

This section illustrates every LLM prompt in details used to construct the CURV Stage 1 and Stage 2 datasets. All prompts target a locally hosted `Qwen/Qwen2.5-7B-Instruct` model exposed as an OpenAI-compatible endpoint, with temperature $\in [0.0, 0.4]$, JSON-mode responses, and ≤ 3 retries on parse failure. Stage 1 prompts (§C.1.1) filter MIMIC-CXR reports and label uncertainty at the sentence and phrase level; Stage 2 prompts (§C.1.2) rewrite uncertainty sentences into hedged form and emit visual-revisit summaries. A summary of the pipeline is given in Tab. 15.

C.1.1 Stage 1 dataset curation prompts

The Stage 1 pipeline turns raw MIMIC-CXR reports into the uncertainty-aware perception dataset $\mathcal{D}_1$ used for Stage 1 SFT. Three LLM call types are used: a report section parser (three header-conditional variants, Tab. 11), a section-level uncertainty extractor (Tab. 12), and a per-phrase uncertainty classifier (Tab. 13).

C.1.2 Stage 2 dataset curation prompts

The Stage 2 pipeline takes each (report sentence, ImaGenome phrase, anatomy, bbox) tuple flagged as uncertain in Stage 1 and asks the LLM to (i) rewrite the sentence so it contains a closed-vocabulary hedge cue and an `<objN>` narrative token, and (ii) emit a short *revisit summary* describing how the corresponding region looked when re-examined. The output of this pass populates the `<thinking>` portion of every Type-A Stage 2 reasoning trajectory. Tab. 14 gives the prompt template (initial call plus the self-correction continuation that fires when the verifier rejects the first output).

C.2 Model Training Prompts

This section lists the exact prompts used in the three training stages of UR^2-MLLM. Stage 1 is a JSON detection task; Stage 2 introduces the free-form structured-report objective with explicit reasoning; Stage 3 applies GRPO on top of Stage 2 without modifying the prompt. A side-by-side summary is given in Tab. 19.

D Uncertainty Analysis

Fig. 6 reports the frequency distribution of uncertainty markers in the full MIMIC-CXR corpus, which contains 276,778 reports and 13.3 M whitespace tokens. We scan every report for occurrences of a closed vocabulary of 35 predefined hedge terms, using case-insensitive whole-word matching with longest-first masking, so that multi-word terms such as *suggestive of* are not double-counted by their constituent single words. In total we observe 155,542 hedge instances, corresponding to 0.56 mentions per report on average, or 1.17% of all word tokens. Bars are sorted in descending order of count. The distribution is heavily long-tailed. The three most frequent markers, namely *likely* (33,645), *may* (23,009), and *appears* (19,243), account for 48.8% of all mentions, while the top ten cover 84% and the bottom ten contribute fewer than 3,000 mentions combined. Multi-word constructions such as *consistent with*, *concerning for*, *compatible with*, and *suggestive of* appear prominently among the top eight, together accounting for 22.3% of all hedge mentions. All 35 terms in the vocabulary appear at least once.

Fig. 7 reports per-thousand-word frequencies for the fourteen most common uncertainty markers across four corpora, namely the MIMIC-CXR ground-truth reports (GT), the Stage 2 SFT model, the full-reward GRPO model, and the no-revisit GRPO ablation. Comparing these distributions reveals how the hedging vocabulary narrows and reshapes as training proceeds.

The ground-truth corpus exhibits a relatively balanced spread of hedging cues. Its five most common markers are *likely* (2.69 per thousand words), *appears* (2.03), *may* (1.63), *consistent with* (1.22), and *could* (0.81), where the most frequent term is only about thirty times more common than the least

	(a) Variant: impression label only — look for unlabelled findings content
system content	You are a radiology report parser. The report below has a labeled `IMPRESSION` section but *no* labeled `FINDINGS` section. Determine whether the report also contains radiographic observations (findings content)—any text that describes the appearance of anatomical structures (e.g. lungs, heart, bones, pleura) using imaging language. Answer *YES* (true) if the report contains any of the following outside the IMPRESSION section: — one or more sentences describing the appearance of anatomical structures; — observations using terms like "clear", "normal", "enlarged", "opacity", "effusion", "no evidence of", "stable", "unchanged", etc.; — a descriptive paragraph providing detail beyond the impression summary. Answer *NO* (false) only if the report contains nothing but the IMPRESSION section and possibly a brief technique / clinical-history note with no imaging observations whatsoever. Reply only with JSON: `{"has_findings": true}` or `{"has_findings": false}`. No explanation.
	(b) Variant: findings label only — look for unlabelled impression / conclusion
system content	You are a radiology report parser. The report below has a labeled `FINDINGS` section but *no* labeled `IMPRESSION` section. Determine whether the report also contains an impression or conclusion—any text that provides a clinical interpretation, summary, or diagnosis based on the findings. Answer *YES* (true) if the report contains any of the following: — one or more sentences summarising the overall clinical interpretation; — a concluding statement giving a diagnosis or differential; — a final paragraph or sentence that reads as a summary / conclusion rather than a detailed observation. Answer *NO* (false) only if the report ends with detailed findings observations and contains no interpretive summary or conclusion anywhere. Reply only with JSON: `{"has_impression": true}` or `{"has_impression": false}`. No explanation.
	(c) Variant: no header labels — look for findings and impression jointly
system content	You are a radiology report parser. The report below has *no* labeled FINDINGS or IMPRESSION sections. Determine whether the report contains *both* types of content, even without labels: A) findings content: text describing the radiographic appearance of anatomical structures (lungs, heart, bones, . . .); B) impression / conclusion content: text that summarises the overall clinical interpretation, diagnosis, or conclusion. Answer *YES* (true) if you can identify both types of content in the report, even if they are not in separate paragraphs. Answer *NO* (false) if the report contains only one type of content, or if it is too brief to clearly contain both. Reply only with JSON: `{"has_both": true}` or `{"has_both": false}`. No explanation.
	User content (shared across all three variants)
user content	`Report:` `{report}`

Table 11: Stage 1A prompt — report section parser (three header-conditional variants). All three variants share the same user content, shown last; `{report}` is replaced with the report body, truncated to 4000 characters. The system content differs by which section headers are present in the raw report.

system content	You are a medical text analysis expert specialising in uncertainty detection. Analyse the provided radiology report section and identify all uncertainty expressions—words or phrases that indicate diagnostic uncertainty, probability, or hedging (e.g. likely, possibly, may, consistent with, cannot exclude, suspicious for).

user content	Analyse the following radiology report section and identify uncertainty expressions.
	`Text: "{text}"`
	Respond with a JSON object containing exactly these keys:
	`"uncertainty_phrases"`: list of uncertainty expressions found (strings);
	`"confidence_level"`: `"high"`, `"medium"`, or `"low"`;
	`"uncertainty_type"`: `"structural"`, `"semantic"`, or `"both"`;
	`"reasoning"`: one-sentence explanation.
	If no uncertainty expressions are found, return
	`{"uncertainty_phrases": [], "confidence_level": "low", "uncertainty_type": "structural", "reasoning": "No uncertainty found."}`

Table 12: Stage 1B prompt — section-level uncertainty extraction. The findings text is truncated to 2000 characters before substitution into `{text}`.

system content	You are a radiology NLP expert. Decide whether a radiology report sentence expresses diagnostic uncertainty about a finding. Uncertainty markers include: *likely, may, might, could, possibly, probably, suspicious, suggestive of, cannot exclude, consistent with, versus, questionable, equivocal, appears, consider, reflect, represent*, etc. Reply only with a JSON object: `{"is_uncertain": true}` or `{"is_uncertain": false}`. No explanation.
user content	`Sentence: "{phrase}"`

Table 13: Stage 1C prompt — per-phrase uncertainty classifier (binary). `{phrase}` is an ImaGenome scene-graph phrase (a noun-phrase or short clause) anchored to a single anatomical bounding box.

frequent one shown and no single cue dominates radiologist phrasing.

This balance does not survive training. All three model variants shift the modal hedge from *likely*, which leads in the GT, to *appears*, which becomes the single most frequent marker in every model corpus, and inflate the rates of a small set of generic cues. Already at the SFT stage, the model uses *appears* at more than twice the GT rate (4.34 against 2.03), *suggestive of* at more than four times the GT rate (2.17 against 0.49), and *possible* at a comparable margin (1.89 against 0.41). The full-reward GRPO model further amplifies the same cues, with *appears* rising to 8.11 and *suggestive of* to 4.74, while the no-revisit ablation reproduces the same pattern at a moderated magnitude (6.65, 2.77, and 2.37 for the three terms). This indicates that the over-reliance on a small group of generic hedges is established during Stage 2 SFT and then further accentuated by the GRPO objective rather than introduced by it.

The revisit reward affects this picture not at the top of the distribution but in its mid-frequency range. For *could*, *possibly*, *possibility of*, *might*, and *potential*, the no-revisit model produces noticeably higher rates than the full-reward GRPO, and in several cases exceeds the GT rate as well, with *could* at 1.51 against 0.81 in the GT, *possibly* at 1.26 against 0.33, and *might* at 0.31 against 0.08. In contrast, the full-reward GRPO concentrates its hedging on the most frequent markers and uses these mid-tail cues sparingly. Removing the revisit reward therefore broadens the hedging vocabulary while leaving the top-end over-use largely intact.

E Reasoning Evaluation

E.1 LLM-based Evaluation Prompts

We score each generated trace on five dimensions (Tab. 20) using GPT-4o as judge. Each dimension receives an integer score from 1 to 5 together with a one-sentence justification; aggregate means are reported in Tab. 21.

E.2 LLM-based Reasoning Evaluation

Tab. 21 reports the GPT-as-judge scores for the `<thinking>` portion of generated reports, averaged over 50 randomly chosen test samples per model. The three model variants differ mainly in which dimension they favor rather than in overall quality. The Stage 2 SFT model achieves the highest clinical plausibility (2.26 against 2.02 for GRPO

(a) Initial call: hedging rewrite + visual-revisit summary (temperature = 0.0)	
system content	You rewrite a single radiology-reasoning sentence and emit a short visual-revisit summary. You obey the schema strictly and never invent findings beyond those in the provided report phrase.

user content	`ORIGINAL_SENTENCE: {sentence}`
	`IMAGENOME_PHRASE: {phrase}`
	`ANATOMY: {anatomy}`
	`OBJ_TOKEN: {obj_token}`
	`REVISIT_STANCE: {stance}`
	Rules.
	1. `hedged_sentence`: rewrite `ORIGINAL_SENTENCE` so it (a) contains *at least one* hedging cue and (b) contains `OBJ_TOKEN` *exactly once*, placed immediately after the `ANATOMY` mention. Preserve clinical meaning. The hedging cue *must* be an exact substring from the closed list (case-insensitive, no inflection—e.g. use "suggests", *not* "suggesting" / "suggest" / "suggested"; use "appears", *not* "appearing"): `[{hedges}]`.
	2. `revisit_summary`: 1–2 sentences, references `OBJ_TOKEN` exactly once, with `stance={stance}` (`confirm` / `qualify` / `alternative`). Do *not* introduce findings absent from `IMAGENOME_PHRASE`. The `revisit_summary` also *must* contain at least one exact hedging cue from the list above —do not inflect.
	Examples of the hedging cue used inline:
	OK: "The cardiac silhouette `{obj_token}` *appears* enlarged . . . "
	OK: "The findings in the left lung `{obj_token}` are *likely* . . . "
	BAD: ". . . *suggesting* pulmonary edema . . . " (inflection)
	BAD: ". . . is *suggestive* . . . " (must be "suggestive of")
	Return strict JSON: `{"hedged_sentence": ". . . ", "revisit_summary": ". . . "}`.

(b) Self-correction turn — fires when (a) output fails the closed-vocabulary hedge-cue verifier (temperature = 0.4, conversation continues from (a))	
user content *(continuation)*	Your previous output did not use a hedge cue from the closed list. Re-emit the JSON. Both fields must use one of: *may, likely, could, appears, suggests, suggestive of, concerning for, consistent with, possible, probable, potential, perhaps*.

Table 14: Stage 2 prompt for hedging rewrite with a self-correction fallback. Part (a) substitutes per-anchor variables for the original sentence, the matched ImaGenome scene-graph phrase, the anatomical region name, a unique narrative tag $\langle \text{obj}_n \rangle$, a stance label chosen from {confirm, qualify, alternative}, and a closed vocabulary of hedge terms (*may*, *likely*, *could*, *appears*, *suggests*, *suggestive of*, *concerning for*, *consistent with*, *possible*, *probable*, *potential*, *perhaps*). Part (b) reuses the same system content as (a). If both passes fail the verifier, a deterministic rule-based fallback inserts "may" or "which may" into the sentence; anchors whose rule-based output still fails are excluded from the SFT corpus.

Stage	Purpose	Input	Output schema
1A	Findings/Impression section parser	full report	`{has_findings|has_impression|has_both: bool}`
1B	Sentence-level uncertainty extraction	report section	`{phrases, conf, type, reason}`
1C	Per-phrase uncertainty classifier	ImaGenome phrase	`{is_uncertain: bool}`
2A	Hedging rewrite + revisit summary	anchor record (5 fields)	`{hedged_sentence, revisit_summary}`
2B	Self-correction (fallback for 2A)	2A output + correction	same as 2A

Table 15: Summary of LLM calls used during data curation. All calls target a locally hosted Qwen2.5-7B-Instruct endpoint with JSON-mode output. Stage 1B is preceded by a regex pre-filter that already removes ∼60% of obviously non-uncertain sentences, so the LLM is invoked only on borderline cases.

system content	You are a radiology AI assistant. Analyze chest X-ray images and identify anatomical regions where the radiologist expressed diagnostic uncertainty.
user content	Analyze this chest X-ray and identify all anatomical regions with diagnostic uncertainty. Return a JSON list where each item contains: `"bbox_2d": [x1, y1, x2, y2]` (normalised coordinates 0–223); `"label":` anatomical region name; `"uncertainty":` the uncertain observation from the report. If no uncertain regions exist, return an empty list `[]`.

assistant target *(example)*	`[`
	`  {`
	`  "bbox_2d": [69, 70, 107, 108],`
	`  "label": "right hilar structures",`
	`  "uncertainty": "The hila are less prominent likely due to . . . "`
	`  }`
	`]`

Table 16: Stage 1 prompt — uncertainty bounding-box detection (SFT). The assistant target is a JSON array (possibly empty) of bounding-box items. Coordinates are normalised to the $[0, 223]$ image grid used by the Stage 1 visual head.

system content	*(none)*
user content	`<image>` Analyze this chest X-ray and generate a structured radiology report with explicit reasoning. Use `<findings>`, `<thinking>`, and `<impression>` sections.

assistant target *(example)*	`<findings>`
	PA chest X-ray shows . . .
	`</findings>`
	`<thinking>`
	. . . natural-language reasoning . . .
	The hazy appearance of the right lower lung field and right hilar lymphadenopathy `<obj1>` suggest pneumonia.
	`detect(`
	`  query="right hilar lymphadenopathy",`
	`  objects=["<obj1><region>[0.46, 0.23, 0.57, 0.44]</region>"])`
	. . . continued reasoning re. `<obj1>` . . .
	`</thinking>`
	`<impression>`
	Bilateral hilar lymphadenopathy and right lower lobe pneumonia . . .
	`</impression>`

Table 17: Stage 2 prompt — reasoning-trace SFT. The assistant target is a three-section report (`<findings>`, `<thinking>`, `<impression>`); inside `<thinking>`, uncertain regions are anchored via `<objN>` narrative tokens and inspected via `detect()` calls whose `<region>` coordinates are normalised to $[0, 1]$. Type-A samples (1,292/1,909) carry at least one revisit anchor; Type-B samples (617/1,909) have no `detect()` calls.

and 1.92 for no-revisit) and the highest faithfulness (2.38 against 2.26 and 2.20), indicating that the SFT objective keeps the model closest to the hedged phrasing of the curated training traces. The

system content	*(none)*
user content *(identical to Stage 2)*	<image> Analyze this chest X-ray and generate a structured radiology report with explicit reasoning. Use <findings>, <thinking>, and <impression> sections.
supervision	GRPO with the four group-relative rewards listed in this table's caption. No token-level target.

Table 18: Stage 3 prompt for GRPO reinforcement learning. The user prompt is identical to Stage 2, while the token-level SFT loss is replaced by four group-relative reward functions. The *format reward* verifies the presence and ordering of <findings>, <thinking>, and <impression>. The *accuracy reward* measures CheXbert-style multi-label agreement with ground-truth findings on the 14 CheXpert labels. The *coherence reward* captures consistency between <thinking> and the <findings>/<impression> sections. The *grounding reward* evaluates the correctness of <objN>, <region>, and detect() triples and their spatial alignment with Stage 1 uncertainty bounding boxes.

	Stage 1	**Stage 2**	**Stage 3**
System prompt	✓	—	—
User prompt template	detection (JSON)	structured report	*same as S2*
Output schema	JSON array	<findings>/<thinking>/<impression>	*same as S2*
detect() in target	—	only Type-A samples	emergent under reward
Supervision	SFT (cross-entropy)	SFT (cross-entropy)	GRPO (4 rewards)

Table 19: Cross-stage prompt and supervision differences. Stages 2 and 3 share an identical prompt, so all behavioral drift between them, such as the detect() trigger rate rising from 25.5% to 51.6%, is attributable to reward shaping rather than prompt engineering.

full-reward GRPO model instead favors uncertainty calibration, scoring slightly higher than the other variants (2.56 against 2.38 for SFT and 2.50 for no-revisit) and echoing the calibration improvement observed in the per-finding analysis above. The no-revisit ablation scores highest on coherence (3.20) and on the findings-to-impression bridge (2.18). Both advantages are partly explained by its longer outputs, a known length bias in LLM-as-judge evaluation. Since all three models cluster within 0.07 of each other on the AVERAGE row, well inside one pooled standard error, the table is best read as a per-dimension diagnostic rather than a global ranking. Taken together, the GRPO objective appears to trade a small amount of faithfulness for better-calibrated hedging, consistent with the per-finding calibration analysis.

E.3 Region-Text Coherence Protocol

For every generated revisit event, we pair the image crop selected by the detect call with the first summary sentence generated after that call. The evaluator scores how specifically the sentence is supported by the visible content of the crop on a five-point scale, where 1 denotes no meaningful correspondence and 5 denotes direct, anatomically and clinically consistent support. We compare three conditions: (i) a same-size random crop paired with the unchanged summary, (ii) the crop selected by the Stage 2 SFT model, and (iii) the crop selected by the full Stage 3 model. The prompt, model version, decoding temperature, and parsing procedure are held fixed across conditions. Aggregate mean scores are reported in Tab. 6. The random-crop condition tests whether the evaluator is responding to generic radiology language rather than to region-specific content.

E.4 Patch-Perturbation Protocol

We sample 200 MIMIC-CXR test cases. At each revisit boundary, we preserve the generated prefix and continue decoding under four embedding conditions: (i) the original selected patch embeddings (clean), (ii) the same number of randomly chosen non-revisited patch embeddings (random-patch control), (iii) zero vectors in place of the selected patch embeddings (zeroed revisit), and (iv) embeddings from an unrelated anatomical region of the same image (mismatched region). All non-intervened inputs and decoding settings are kept fixed. We compute CheXbert Macro-F1 on the completed reports and define "Changed" as the percentage of cases whose regenerated post-revisit summary differs from its clean-decoding counterpart. The random-patch control distinguishes disruption caused by the selected evidence from dis-

ruption caused by modifying an equal number of arbitrary visual tokens. Results are reported in Tab. 7.

F Case Study

To better understand the behavior of UR^2-MLLM, we present two qualitative case studies from MIMIC-CXR in Fig. 8, illustrating both a successful prediction and a more challenging failure scenario.

Fig. 8(a) shows a successful case where the model correctly handles a clinically subtle finding. The ground-truth report notes that low lung volumes exaggerate the cardiomediastinal contours while the heart size remains top normal. Our model reaches the same conclusion through an explicit reasoning trace. It first detects the right lung and the left lung as separate regions, grounds them via bounding box predictions, and then evaluates the cardiac silhouette within the context of the surrounding lung fields. By revisiting these anatomical regions before issuing a final judgment, the model avoids the common pitfall of over-pathologizing a low-volume study and produces a generated report that closely aligns with the ground truth on heart size, pneumothorax, and pleural effusion. The predicted bounding boxes also overlap well with the ImaGenome ground-truth regions, indicating that the visual grounding is consistent with the textual reasoning.

Fig. 8(b) presents a more challenging case in which the underlying pathology is severe and involves multiple coexisting findings. The ground-truth report describes extremely low lung volumes, diffuse interstitial opacities, and superimposed pulmonary edema on a background of pulmonary fibrosis, with cardiomediastinal and hilar contours remaining within normal limits. Our model correctly identifies the hyperinflation pattern and grounds both the cardiac silhouette and the right hemidiaphragm through its reasoning process, yet it ultimately commits to a differential diagnosis of heart failure with pulmonary edema and reports cardiomegaly together with a moderate right pleural effusion that the radiologist does not confirm. Although the final output is incorrect, the reasoning trace itself remains valuable. It reveals exactly where the error originates, namely at the transition from regional findings to the global differential diagnosis, where the model over-commits to a familiar disease pattern instead of staying faithful to the visually grounded evidence. This kind of transparency is difficult to obtain from end-to-end models and points to a clear direction for future improvement, which is to introduce stronger consistency constraints between the grounded regions and the final diagnostic conclusion.

G Licenses for External Assets

This appendix summarizes the sources, URLs, licenses, and access conditions of the publicly available datasets and pre-trained models used in this work. Details can be found in Tab. 22.

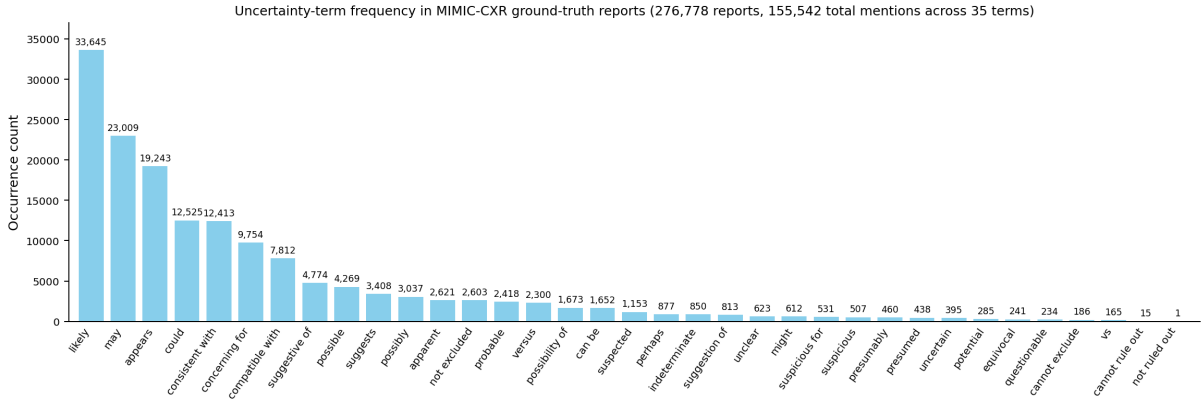


Figure 6: Frequency distribution of uncertainty markers in MIMIC-CXR ground-truth reports. All 276,778 reports (train+val+test) were scanned for occurrences of the 35-term closed vocabulary used during Stage 2 trace synthesis (longest-first masked matching, case-insensitive, whole-word boundary). Top three markers (*likely*, *may*, *appears*) cover 48.8% of all 155,542 mentions; the top ten cover 84%. Multi-word constructions (*consistent with*, *concerning for*, *compatible with*) appear prominently among the top seven.

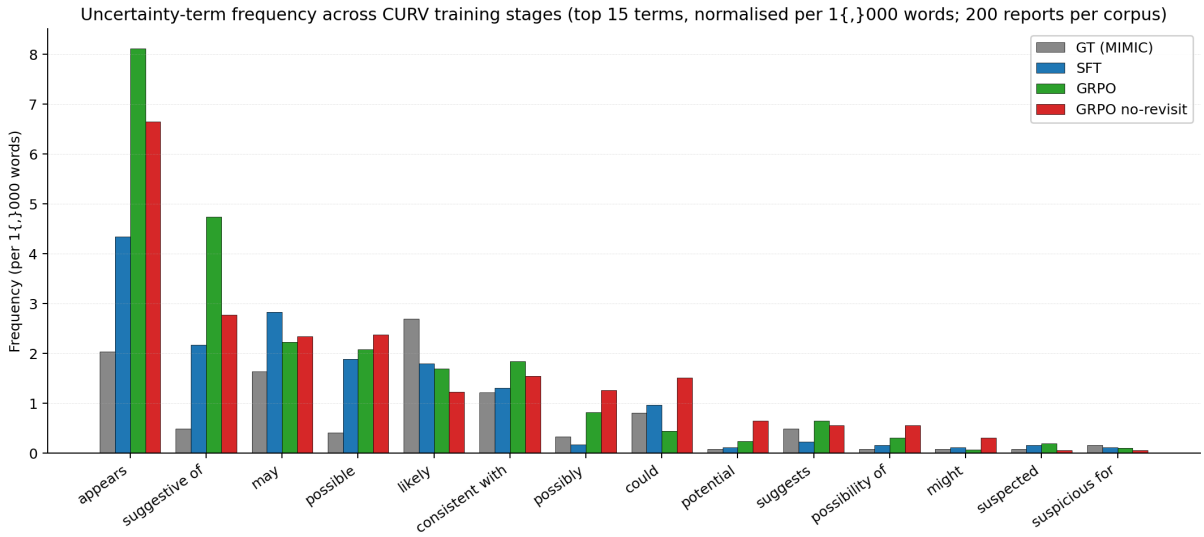


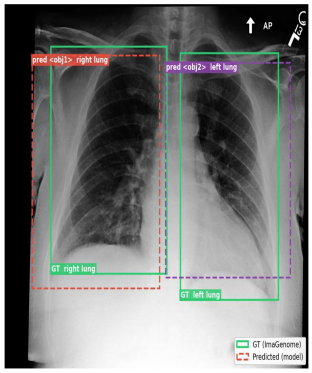
Figure 7: Per-thousand-word frequency of the top fourteen uncertainty markers across four corpora. Bars within each group correspond, from left to right, to ground-truth MIMIC-CXR reports, the Stage 2 SFT model, the full-reward GRPO model, and the no-revisit GRPO ablation. All four corpora are matched on the same 200-report test split, with rates normalised by total word count.

system content	You are a senior radiologist evaluating an AI-generated chest X-ray reasoning trace. You are given the ground-truth (radiologist-written) report and an AI-generated output that contains `<findings>`, `<thinking>`, and `<impression>` sections. You will score the AI output on five dimensions on a 1–5 5-point rating scale. Output strict JSON only. Length and writing style are *not* criteria. You do not know which AI model produced the output.
	User content (scoring rubric, 1–5 integer per dimension; higher is better)
1. clinical plausibility	How medically sensible is the reasoning given a chest X-ray? 1 = medically incorrect or nonsensical. 3 = mostly plausible with notable inaccuracies. 5 = fully clinically valid; a radiologist would accept the inferences.
2. faithfulness	Compare claims in the AI output (especially in `<findings>` and `<thinking>`) against the ground-truth report. 1 = severely hallucinated; the trace asserts findings clearly absent from the GT. 3 = partially grounded; some unsupported claims. 5 = fully grounded; no findings or locations introduced beyond what the GT supports.
3. coherence	Internal logical consistency of `<thinking>`. 1 = self-contradictory, jumps between unrelated claims. 3 = mostly coherent, with some logical gaps. 5 = clean logical progression from observation to inference, no contradictions.
4. findings→impression bridge	How well does `<thinking>` motivate `<impression>`? 1 = impression conclusions are not supported by `<thinking>`. 3 = some impression statements are supported, others are not. 5 = every claim in `<impression>` is traceable to reasoning in `<thinking>`.
5. uncertainty calibration	Are hedge words (*likely, may, appears, suggestive of, possible, consistent with*, ...) and confident assertions placed appropriately? 1 = hedges used randomly or missing where evidence is weak. 3 = some hedges are warranted, others misplaced. 5 = hedges only where evidence is genuinely uncertain; commitment where evidence is strong.
	User content (input substitution and required output)

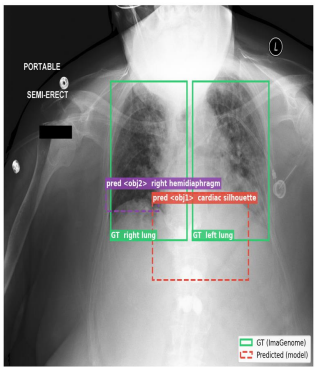
inputs	`=== GROUND-TRUTH REPORT (written by a radiologist) ===`
	`{gt_report}`
	
	`=== AI-GENERATED OUTPUT (contains <findings>, <thinking>, <impression>) ===`
	`{generated}`

output schema	Return a single JSON object (no markdown, no trailing text):
	`{`
	`"clinical_plausibility": {"score": <1-5>, "justification": ". . . "},`
	`"faithfulness": {"score": <1-5>, "justification": ". . . "},`
	`"coherence": {"score": <1-5>, "justification": ". . . "},`
	`"bridge": {"score": <1-5>, "justification": ". . . "},`
	`"uncertainty_calibration": {"score": <1-5>, "justification": ". . . "}`
	`}`

Table 20: LLM-as-judge prompt used to score the `<thinking>` portion of generated reasoning traces. The judge is GPT-4o (`gpt-4o-2024-11-20`) called with temperature $= 0$ and JSON-mode output; identical `seed` is used across all reports to ensure reproducibility. For every test report and every model under evaluation, the prompt is instantiated by substituting the radiologist-written ground-truth report into `{gt_report}` and the model's full `<findings>`/`<thinking>`/`<impression>` generation into `{generated}`. All five dimensions are scored on a 1–5 Likert scale; each score is accompanied by a one-sentence justification. Aggregate results are reported in Tab. 21.

Generated / GT Matched Reports	UR^2-MLLM's Reasoning Process	Generated / GT Revisit Region
GT Report: Low lung volumes exaggerate the cardiomediastinal contours, however the heart size is top normal. There is mild bibasilar atelectasis. No focal consolidations concerning for pneumonia are identified. There is no pleural effusion or pneumothorax. The visualized osseous structures are unremarkable. ***Generated Report :*** The heart is in normal size. The right lung shows no pleural effusion. No pneumothorax is identified. No pleural masses or other focal abnormalities are noted. Minimal right pleural effusion. Slight enlargement of the heart, suggestive of mild cardiomegaly.	In analyzing this chest X-ray, my approach begins with a systematic evaluation of the provided findings... a slightly enlarged heart <obj1>, likely no pleural effusion in the right lung, no pneumothorax...**detect(query="right lung", objects=["<obj1><region>[0.09, 0.14, 0.51, 0.78]<|copy_9|><|copy_10|><|copy_11|><|copy_26|><|copy_19|><|copy_28|><|copy_38|><|copy_28|><|copy_25|><|copy_35|><|copy_34|><|copy_36|><|copy_43|><|copy_45|><|copy_38|><|copy_44|><|copy_42|><|copy_49|><|copy_42|></region>"])**The findings in the right lung <obj1> suggest that there is minimal pleural effusion... Additionally, the heart <obj2> appears normal. **detect(query="left lung", objects=["<obj2><region>[0.53, 0.16, 0.94, 0.75]<|copy_6|><|copy_14|><|copy_5|><|copy_13|><|copy_21|><|copy_20|><|copy_29|><|copy_30|><|copy_22|><|copy_38|><|copy_50|><|copy_52|><|copy_28|><|copy_33|><|copy_35|><|copy_37|></region>"])** The findings in the left lung <obj2> suggest that the heart is in normal size. First, I consider the finding of a slightly enlarged heart... …	

(a) Success sample case.

Generated / GT Matched Reports	UR^2-MLLM's Reasoning Process	Generated / GT Revisit Region
GT Report: **…** **Lung volumes are extremely low** resulting in bronchovascular crowding and limited evaluation of the lung bases. Diffuse interstitial opacities have increased, and despite the low lung volumes, findings are consistent with superimposed pulmonary edema on a background of pulmonary fibrosis. **No large pleural effusion is evident.** There is no pneumothorax. **Cardiomediastinal and hilar contours are within normal limits.** ***Generated Report :*** The lungs are **hyperinflated**. The heart is **enlarged**, but the pulmonary vasculature remains within normal limits. There is a **moderate right pleural effusion** with **mild elevation of the right hemidiaphragm.** No focal consolidation or other evidence of pneumonia is seen. Hyperinflation, cardiomegaly, moderate right pleural effusion, …	The first notable observation is the **hyperinflation** of the lungs... Hyperinflation can result from chronic restrictive lung disease... Next, I note the **enlarged cardiac silhouette** <obj1>, which indicates cardiomegaly. **detect(query="cardiac silhouette", objects=["<obj1><region>[0.48, 0.52, 0.79, 0.76]<|copy_27|><|copy_23|><|copy_35|><|copy_45|><|copy_52|><|copy_38|><|copy_45|><|copy_46|></region>"])** The presence of a **moderate right pleural effusion** with mild elevation of the right hemidiaphragm <obj2> is significant. **detect(query="right hemidiaphragm", objects=["<obj2><region>[0.33, 0.48, 0.50, 0.57]<|copy_29|></region>"])** [Extended differential diagnosis follows — concludes **"heart failure with pulmonary edema"**]	

(b) Failure sample case.

Figure 8: Two qualitative case studies from the MIMIC-CXR.

Dimension	**SFT**	**GRPO**	**GRPO no-revisit**
Clinical plausibility	2.26 ± 0.60	2.02 ± 0.59	1.92 ± 0.67
Faithfulness	2.38 ± 0.53	2.26 ± 0.44	2.20 ± 0.45
Coherence	3.06 ± 0.74	3.00 ± 0.53	3.20 ± 0.40
Findings→Impression bridge	2.16 ± 0.74	2.06 ± 0.55	2.18 ± 0.52
Uncertainty calibration	2.38 ± 0.67	2.56 ± 0.64	2.50 ± 0.65
Average	2.45	2.38	2.40

Table 21: Mean $\pm$ std of GPT-4o-as-judge scores (1 to 5) on the `<thinking>` portion ($n = 50$ per model). "Bridge" denotes how well `<thinking>` supports `<impression>`. Best in bold; SE of the mean ≤ 0.11. Full prompt in Tab. 20.

Datasets

- **MIMIC-CXR (version 2.1.0):**
  - **Source:** Johnson *et al.* (Johnson et al., 2024).
  - **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.* (Wu et al., 2021).
  - **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 UR^2-MLLM):**
  - **Source:** Bai *et al.* (Team, 2025b).
  - **URL:** https://github.com/QwenLM/Qwen3
  - **License:** Apache 2.0 License.
- **LLaVA-1.5-7B:**
  - **Source:** Liu *et al.* (Liu et al., 2024a).
  - **URL:** https://github.com/haotian-liu/LLaVA
  - **License:** Apache 2.0 License.
- **MAIRA-2:**
  - **Source:** Bannur *et al.* (Bannur et al., 2024).
  - **URL:** https://huggingface.co/microsoft/maira-2
  - **License:** Microsoft Research License Terms.
- **HuatuoGPT-Vision-7B:**
  - **Source:** Chen *et al.* (Chen et al., 2024).
  - **URL:** https://huggingface.co/FreedomIntelligence/HuatuoGPT-Vision-7B
  - **License:** Apache 2.0 License.
- **RadGraph:**
  - **Source:** Jain *et al.* (Jain et al., 2021).
  - **URL:** https://huggingface.co/StanfordAIMI/RRG_scorers
  - **License:** MIT License.
- **CheXbert:**
  - **Source:** Smit *et al.* (Smit et al., 2020).
  - **URL:** https://huggingface.co/StanfordAIMI/RRG_scorers
  - **License:** MIT License.

Table 22: Licenses and access conditions of the publicly available datasets and pre-trained models used in this work.