REVA-PO: Stabilizing Reinforcement Learning for Chest X-ray Report Generation

Li Guo, Anas M. Tahir, and Z. Jane Wang

University of British Columbia, Vancouver, BC, Canada
lguo@ece.ubc.ca

Abstract. Automated chest X-ray report generation has recently benefited from reinforcement learning (RL) and large language models. However, RL training often suffers from instability or limited exploration due to fixed Kullback-Leibler (KL) regularization and a static reference policy that accumulates KL pressure over time. We propose Response-Weighted and Validation-Anchored Policy Optimization (REVA-PO), a RL framework that stabilizes long-term training via Response-Weighted Regularization (RER) and Validation-Anchored Policy Reset (VAPR). RER dynamically adjusts per-response KL weights based on advantage and reference-policy entropy, relaxing constraints for high-quality responses while tightening them for low-quality ones. Complementarily, VAPR periodically synchronizes the reference and current policies to the best validation checkpoint, resetting accumulated regularization pressure to expand the viable exploration space. To ensure a robust starting point, we employ a three-stage pipeline consisting of warm-up training, classifier-guided supervised fine-tuning, and RL. Extensive evaluations on MIMIC-CXR and IU-Xray demonstrate that REVA-PO sets new state-of-the-art benchmarks in both linguistic quality and clinical accuracy. Notably, BLEU-4 improves by 5.1% on MIMIC-CXR and 3.6% on IU-Xray, while CheXpert F1 and RadGraph F1 scores increase by 4.5% and 12.8%, respectively, over prior leading methods. The code is publicly available at https://github.com/LiGuo12/REVA_PO/.

Keywords: Reinforcement Learning · KL Regularization · Chest X-ray Report Generation

1 Introduction

Chest X-rays (CXR) are essential for diagnosing and monitoring chest organ conditions, yet the demand for interpretation often exceeds radiologist capacity, increasing turnaround times and diagnostic risks [4, 22, 36]. While automated CXR report generation (CXR-RG) has progressed significantly [13, 41], generated reports still struggle with nuanced word choice and sentence fluency compared to professional reports.

To improve report quality, recent work has integrated large language models (LLMs) [18,26,43,45,58]. Standard approaches often utilize contrastive learning to

map radiographic features into LLM embedding space—a process that is memory-intensive and prone to overfitting on limited medical datasets. While the emerging large vision-language models (LVLMs) pretrained on broad image–text data already align visual and textual spaces [1, 11, 53]. Inspired by the advantages of LVLMs, we propose adopting a pretrained LVLM for CXR-RG, which eliminates the need to train a separate image encoder and reduces memory use while mitigating overfitting.

From a training perspective, reinforcement learning from human feedback (RLHF) has demonstrated superior performance in text generation [2, 30, 67]. Standard proximal policy optimization (PPO) [37] requires a critic model often as large as the policy itself. To optimize resource use, we adopt group relative policy optimization (GRPO) [12, 38], which omits the critic by estimating baselines from multiple responses per prompt. Furthermore, training a reward model requires labeled human-preference data, which is costly to collect from radiologists. Following previous RL work in CXR-RG [20, 32], we utilize the BLEU-4 score between the generated and reference reports as the reward, thereby avoiding the need for a learned reward model.

While Kullback–Leibler (KL) regularization is the standard safeguard against RL instability, a fixed KL coefficient β often creates a "lose-lose" scenario: it either stifles productive exploration or fails to prevent destabilizing updates. To address this, recent work has explored adaptive β scheduling at the batch level [37, 51] or finer-grained adjustments at the instance level [23, 65]. However, batch-level adjustments often mask harmful individual trajectories by averaging across the batch. Alternatively, instance-level methods either inflate computational overhead with an auxiliary critic [65] or remain strictly dependent on the availability of preference data [23].

To address the above limitations, we propose Response-Weighted Regularization (RER) for fine-grained, response-level control, which is conditioned on both output quality and reference-policy entropy. As a seamless, drop-in replacement for PPO or GRPO objectives, RER requires no auxiliary components or costly preference data. We define a per-response regularization weight β_i as a function of the advantage A_i and reference entropy $\mathcal{H}_i^{\text{ref}}$. Specifically, for beneficial exploration ($A_i > 0$), β_i is decreased as reference entropy $\mathcal{H}_i^{\text{ref}}$ increases, thereby shielding novel, high-performing behaviors from excessive penalization. Conversely, for harmful deviations ($A_i < 0$), a high $\mathcal{H}_i^{\text{ref}}$ triggers an increase in β_i to pull the current policy back toward the reference distribution. By strategically relaxing constraints on productive discovery while tightening them on poor trajectories, RER simultaneously enhances sample efficiency and training stability.

Policy collapse can also occur as the reference policy becomes outdated, causing KL loss to dominate the gradient. While some researchers employ an exponential moving average (EMA) to allow the reference policy to track the current policy [6,34], EMA is inherently vulnerable to reward hacking. If the policy begins to exploit a flawed reward signal, the EMA incorporates this suboptimal behavior into the reference, creating a vicious cycle where future penalties become

ineffective at correcting the drift. To resolve these issues, we propose Validation-Anchored Policy Reset (VAPR), a mechanism that resets both the reference policy and current policy to the validation-best checkpoint at fixed intervals. This ensures that every reference update is grounded in empirical validity rather than blind tracking. Unlike the inherent lag in EMA, VAPR's simultaneous update of the reference and current policies nullifies accumulated regularization pressure, providing a fresh, stabilized baseline for continued optimization.

Collectively, RER and VAPR form our Response-Weighted and Validation-Anchored Policy Optimization (REVA-PO) framework. By balancing response-level flexibility with long-term structural stability, REVA-PO enhances the efficiency and quality of exploration while effectively insulating the training process against catastrophic policy collapse.

RL performance is also fundamentally constrained by the quality of the initial policy. A poor initial policy can reinforce stochastic errors rather than meaningful clinical insights. To ensure a robust starting point, we employ a three-stage training pipeline: (1) warm-up, (2) classifier-guided SFT, and (3) RL with REVA-PO. This staged approach ensures that the LVLM begins the RL phase with high-quality exploration and a reliable reference for KL regularization.

Our contributions are summarized as follows:

- **Response-Weighted Regularization (RER)**: We propose RER, a response-level mechanism that dynamically scales β_i for each candidate based on its advantage and reference entropy, protecting high-quality discovery while penalizing unstable updates.
- **Validation-Anchored Policy Reset (VAPR)**: We propose VAPR to prevent training collapse by periodically resetting the reference and current policies to the best validation checkpoint, clearing accumulated regularization pressure.
- **Efficient LVLM-based CXR-RG**: We introduce a three-stage pipeline that leverages a pretrained LVLM to eliminate the need for training a separate image encoder, thereby reducing memory overhead and mitigating overfitting risks on smaller datasets like IU-Xray.
- **State-of-the-Art Performance**: Extensive evaluations on MIMIC-CXR and IU-Xray validate the efficacy of REVA-PO. Our method achieves relative BLEU-4 gains of 5.1% and 3.6%, respectively, while attaining peak CheXpert F1 (0.506) and RadGraph F1 (0.246) scores. These results confirm superior linguistic fluency and clinical accuracy.

2 Related Work

2.1 Chest X-ray Report Generation

Prior work on CXR-RG spans joint classification, knowledge graphs, RL, and LLMs; each has trade-offs. Adding a disease classification head and training with a joint loss can raise lesion sensitivity [19,44,47,49,56], but pushes features toward coarse categories, which hurts descriptive detail. We instead use a frozen classifier to supply disease cues without backpropagating gradients into the generator.

Medical knowledge graphs can inject prior knowledge [16, 24, 27, 57, 63], but are costly to build and maintain. Causal-intervention methods, such as CMCRL [5], aim to mitigate spurious vision-language correlations and achieve promising results.

RL has been adopted to optimize sequence-level metrics for CXR-RG [32, 49, 50]. Most prior approaches rely on self-critical sequence training (SCST) [35], a policy-gradient method with a self-critical baseline for non-differentiable metrics. However, because SCST samples only a single response per prompt, it suffers from high-variance gradients and restricted exploration, which has limited the performance ceiling of RL-based CXR-RG methods. To overcome this, our REVA-PO framework builds upon GRPO by sampling a group of responses per prompt, thereby reducing gradient variance and promoting exploration.

Recently, LLM-based methods have significantly improved report quality [18, 26, 43, 45, 46, 58]. Most of these approaches rely on contrastive learning to align radiographs with LLMs, necessitating large batch sizes and substantial GPU memory. In contrast, we utilize a LVLM [1] that is pre-aligned on general image–text data. This approach bypasses the need for contrastive pretraining and effectively reduces memory overhead.

2.2 Reinforcement Learning

As LVLMs are still emerging, research into LVLM-specific RL remains limited [61,62]. However, due to their architectural similarities to LLMs, many established RL methods transfer effectively to the multimodal domain [39, 60]. In traditional RLHF, PPO [37] optimizes a clipped objective using a critic and a reward model trained on human preferences. However, this pipeline requires substantial compute and annotation overhead [2, 30, 67]. Direct Preference Optimization (DPO) simplifies this by replacing the reward model with a log-ratio objective between current and reference policies, yet it still necessitates preference pairs [33]. In contrast, GRPO eliminates the critic and preference pairs by sampling multiple responses per prompt and using group-average rewards as a baseline [12, 38]. Despite these advantages, GRPO relies on a fixed KL β that ensures stability at the cost of restricted exploration.

Recent work has explored dynamic β adjustment to address this limitation. PPO [37] adapts β based on batch average KL divergence, increasing the penalty when divergence exceeds a target to maintain stability, which can inadvertently throttle learning. β-DPO [51] and ε-DPO [23] adapt β using signals derived from preference pairs, which remain dependent on paired data. While Zheng et al. [65] use a critic-based classifier to scale β for high performing samples, this significantly increases memory costs. Our proposed RER provides fine-grained control for each response without requiring preference data or additional components. Furthermore, a static reference policy allows KL pressure to accumulate, leading to sudden training collapse. While some methods use EMA to update the reference [6,34], blind tracking risks absorbing erroneous behaviors and inherently lags behind the current policy. We instead propose VAPR, which periodically

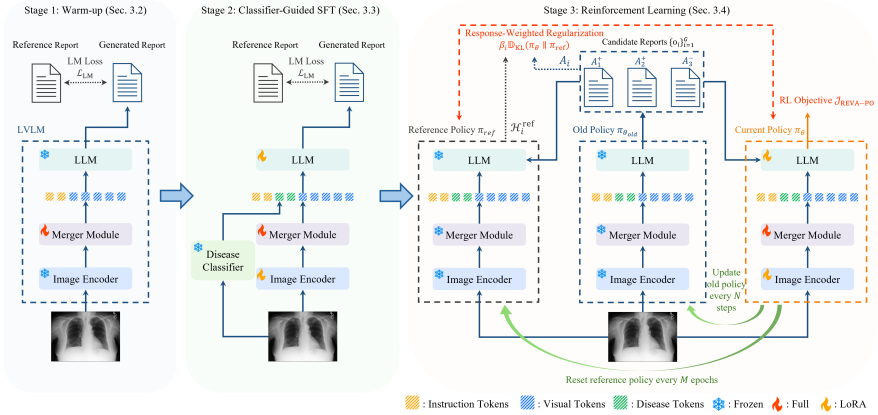


Fig. 1: The three-stage training pipeline. *Stage 1: Warm-up.* Freeze the image encoder and LLM; train only the merger with $\mathcal{L}_{\text{LM}}$. *Stage 2: Classifier-guided SFT.* Build prompt $q = [I, V, D]$ by appending predicted labels D; freeze the classifier; train the merger and LoRA-tune the encoder and LLM with $\mathcal{L}_{\text{LM}}$. *Stage 3: RL with REVA-PO.* $\pi_{\theta_{\text{old}}}$ samples G reports, computes advantages A_i, and maximizes $\mathcal{J}_{\text{REVA-PO}}$ with entropy bonus $\mathcal{H}_{i,t}^{\text{cur}}$ and response-weighted regularization $\beta_i(A_i, \mathcal{H}_i^{\text{ref}})$; reset π_{ref} every M epochs and synchronize $\pi_{\theta_{\text{old}}}$ every N steps. Classifier omitted in Stage 3. The legend indicates instruction, visual, and disease tokens, and the training status of each module (frozen, full, LoRA).

realigns the reference and current policies. This resets the effective KL penalty and ensures long-term training stability.

3 Method

Overview. We introduce a three-stage pipeline (Fig. 1): (i) a warm-up phase that trains only the merger to adapt the LVLM to radiographs (Sec. 3.1); (ii) classifier-guided SFT, which injects disease-specific cues (Sec. 3.2); and (iii) RL with REVA-PO, which leverages RER to scale KL weights for each response and VAPR to maintain long-term stability while encouraging exploration (Sec. 3.3).

3.1 Warm-up

An LVLM consists of an image encoder, a merger that maps visual features to the LLM embedding space, and the LLM. Inspired by [28, 66], we freeze the image encoder and the LLM and train only the merger. This step aligns the scale and distribution of mapped visual features with the radiological domain while keeping the pretrained representations unchanged. With both endpoints fixed, optimization remains stable and the frozen components do not drift.

Given a prompt $q = [I, V]$ with instruction tokens I and visual tokens V, and reference tokens $y_{1:T}$, the language modeling loss is

$$\mathcal{L}_{\mathrm{LM}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta\big(y_t \mid y_{<t}, q\big). \tag{1}$$

3.2 Classifier-Guided Supervised Fine-Tuning

After warm-up, the LVLM can produce reports that resemble reference reports in vocabulary and sentence structure, yet its disease-level diagnostic accuracy is still limited. Drawing inspiration from PromptMRG [19], we integrate a pretrained disease classifier to provide explicit clinical cues. Given a CXR, the classifier predicts 14 disease categories following the CheXpert [17] taxonomy. We append the predicted positive labels to the prompt to guide the LLM toward lesion-relevant visual tokens.

Concretely, if the classifier predicts "Cardiomegaly" and "Lung Opacity" (denote these as D), we form $q = [I, D, V]$, where I are instruction tokens and V are visual tokens. The labels D guide attention to the corresponding anatomical regions and reduce missed findings.

Since the model is warmed up, we increase the number of trainable parameters. The merger is fully trainable, all linear layers of the image encoder and the LLM are fine-tuned with LoRA [14], and the classifier is frozen. Training continues with $\mathcal{L}_{\mathrm{LM}}$ in Eq. (1).

3.3 Reinforcement Learning

Stage 3 is initialized from the Stage 2 weights and uses the same trainable parameters. Given an input image with prompt q and its reference report a, the old policy $\pi_{\theta_{\mathrm{old}}}$ samples a group of G candidate reports $\{o_i\}_{i=1}^{G}$. Each o_i receives a reward R_i given by the BLEU-4 score between o_i and a. As illustrated in Supplementary Material A, we use BLEU-4 as a proxy reward for its low variance and strong directional alignment with clinical efficacy metrics. Subsequently, we compute a group-relative standardized advantage

$$A_i \;=\; \frac{R_i - \mathrm{mean}\big(\{R_j\}_{j=1}^{G}\big)}{\mathrm{std}\big(\{R_j\}_{j=1}^{G}\big)}. \tag{2}$$

REVA-PO Objective. We optimize the policy π_θ using a GRPO-style objective, which incorporates a clipped surrogate loss, the proposed RER mechanism β_i, and an entropy bonus. The objective is

$$\mathcal{J}_{\mathrm{REVA-PO}}(\theta) = \mathbb{E}[(q,a)\sim\mathcal{D},\ \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot\mid q)] \left[\frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\Big(\ell_{i,t}(\theta) - \beta_i\,\mathbb{D}_{\mathrm{KL}}(\pi_\theta\,\|\,\pi_{\mathrm{ref}})\Big)\right], \tag{3}$$

where the clipped term is

$$\ell_{i,t}(\theta) = \min\Big(r_{i,t}(\theta)\, A_i,\ \mathrm{clip}(r_{i,t}(\theta),\, 1-\epsilon_{\mathrm{low}},\, 1+\epsilon_{\mathrm{high}})\, A_i\Big), \tag{4}$$

with importance ratio

$$r_{i,t}(\theta) \ = \ \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{5}$$

The Kullback-Leibler (KL) divergence uses the unbiased estimator

$$\begin{aligned}\mathbb{D}_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}}) = {} & \frac{\pi_{\mathrm{ref}}(o_{i,t} \mid q, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})} \\ & - \log \frac{\pi_{\mathrm{ref}}(o_{i,t} \mid q, o_{i,<t})}{\pi_\theta(o_{i,t} \mid q, o_{i,<t})} - 1.\end{aligned} \tag{6}$$

As suggested by [59], we use the token-level average $\frac{1}{\sum_i |o_i|} \sum_i \sum_t (\cdot)$ in Eq. (3) to give each token equal weight and avoid downweighting long responses. Maximizing $\mathcal{J}_{\mathrm{REVA-PO}}$ increases the probability of high-advantage responses and decreases that of low-advantage responses.

Response-Weighted Regularization. KL regularization stabilizes learning by penalizing deviation from the reference policy π_{ref}. However, a fixed weight β can over-penalize beneficial exploration and under-penalize harmful drifts. We therefore introduce RER, which assigns a response-level coefficient β_i to each candidate o_i based on its advantage A_i and the reference-policy entropy along the sequence. The sequence-level reference-policy entropy is defined as

$$\mathcal{H}_i^{\mathrm{ref}} = -\frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \sum_{v \in \mathcal{V}} \pi_{\mathrm{ref}}(v \mid q, o_{i,<t}) \, \log \pi_{\mathrm{ref}}(v \mid q, o_{i,<t}), \tag{7}$$

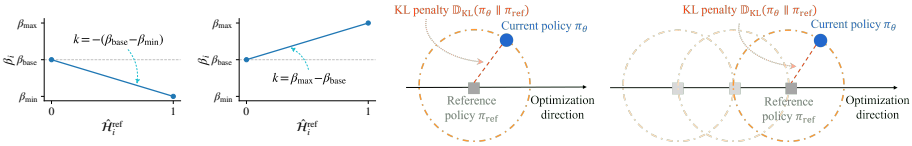
where $\mathcal{V}$ is the token vocabulary. Using the token-normalized entropy makes $\mathcal{H}_i^{\mathrm{ref}}$ insensitive to response length. We normalize this entropy within the group to obtain

$$\hat{\mathcal{H}}_i^{\mathrm{ref}} \ = \ \frac{\mathcal{H}_i^{\mathrm{ref}} - \min\big(\{\mathcal{H}_j^{\mathrm{ref}}\}_{j=1}^G\big)}{\max\big(\{\mathcal{H}_j^{\mathrm{ref}}\}_{j=1}^G\big) - \min\big(\{\mathcal{H}_j^{\mathrm{ref}}\}_{j=1}^G\big)} \ \in \ [0,1]. \tag{8}$$

We then set β_i as a function of A_i and $\hat{\mathcal{H}}_i^{\mathrm{ref}}$

$$\beta_i = \begin{cases} \beta_{\min} + (\beta_{\mathrm{base}} - \beta_{\min})\,(1 - \hat{\mathcal{H}}_i^{\mathrm{ref}}), & A_i > 0, \\ \beta_{\mathrm{base}} + (\beta_{\max} - \beta_{\mathrm{base}})\,\hat{\mathcal{H}}_i^{\mathrm{ref}}, & A_i < 0, \\ \beta_{\mathrm{base}}, & A_i = 0, \end{cases} \tag{9}$$

with $\beta_{\min} \le \beta_{\mathrm{base}} \le \beta_{\max}$.



(a) $A_i > 0$ **(b)** $A_i < 0$ **(c)** Fixed Reference **(d)** VAPR

Fig. 2: (Left) Mapping from the normalized reference-policy entropy $\hat{\mathcal{H}}_i^{\text{ref}}$ to the per-response KL weight β_i. For advantage $A_i > 0$, the line has slope $k = -(\beta_{\text{base}} - \beta_{\min})$; for $A_i < 0$, the line has slope $k = \beta_{\max} - \beta_{\text{base}}$. (Right) Fixed Reference vs. Validation-Anchored Policy Reset (VAPR). The reference policy π_{ref} is the anchor (gray), the current policy π_θ is the ball (blue), and the KL penalty $\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$ is the elastic string. With fixed π_{ref}, the KL term keeps exploration local; VAPR resets π_{ref} and π_θ to the validation-best checkpoint, effectively releasing the regularization pressure to expand the reachable region.

(i) For $A_i > 0$ (better-than-mean response), a larger $\hat{\mathcal{H}}_i^{\text{ref}}$ indicates the reference policy is uncertain about producing this good response. We therefore reduce β_i toward $\beta_{\min}$ to avoid constraining promising updates.

(ii) For $A_i < 0$ (worse-than-mean response), a larger $\hat{\mathcal{H}}_i^{\text{ref}}$ implies the reference policy is less likely to generate such a poor response. We therefore increase β_i toward $\beta_{\max}$, which strengthens the pull that drives π_θ back toward π_{ref}.

The mapping $(A_i, \hat{\mathcal{H}}_i^{\text{ref}}) \mapsto \beta_i$ is bounded and piecewise monotonic with respect to $\hat{\mathcal{H}}_i^{\text{ref}}$ (Figs. 2a and 2b), preserving the numerical stability of β_i throughout the training process.

Validation-Anchored Policy Reset. Although RER relaxes penalties on promising exploratory actions, exploration remains fundamentally constrained by $\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$ when π_{ref} is fixed. As π_θ drifts from π_{ref}, the growing KL term pulls the policy back, restricting exploration to a static trust region around the initial reference.

To overcome this, we introduce VAPR, which periodically expands the explorable space while maintaining stability. After each epoch, we evaluate π_θ on the validation set and save the checkpoint θ_{best} that yields the highest validation score s_{val}:

$$s_{\text{val}} = \frac{1}{|\mathcal{M}| + 1} \left(\sum_{m \in \mathcal{M}} m + \lambda \cdot \text{RG} \right), \tag{10}$$

where $\mathcal{M}$ is the set of natural language generation metrics, RG denotes the RadGraph F1 [9] as a proxy for clinical factual consistency, and λ is a balancing coefficient.

Every M epochs, we reset π_{ref} and synchronize the $\pi_{\theta_{\text{old}}}$ and π_θ to θ_{best}. This reset enforces $\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) = 0$, allowing optimization to proceed from the new anchor without accumulated regularization pressure. Synchronizing $\pi_{\theta_{\text{old}}}$ simultaneously mitigates off-policy sampling mismatch. Algorithm 1 outlines the full procedure.

Algorithm 1 REVA-PO: Response-Weighted and Validation-Anchored Policy Optimization

Input: policy π_θ; dataset $\mathcal{D}$; group size G; learning rate η; sampling synchronization period N (steps); reference reset period M (epochs); total epochs E; steps per epoch S

$\pi_{\text{ref}} \leftarrow \pi_\theta, \quad \pi_{\theta_{\text{old}}} \leftarrow \pi_\theta, \quad \theta_{\text{best}} \leftarrow \theta$
for epoch $= 1$ **to** E **do**
  for step $= 1$ **to** S **do**
    sample $(q, a) \sim \mathcal{D}$
    sample G candidates $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q)$
    compute rewards $\{R_i\}_{i=1}^{G}$ for the sampled $\{o_i\}$
    compute $\mathcal{J}_{\text{REVA-PO}}(\theta; \{o_i\}, \pi_{\text{ref}})$ ▷ Eq. (3)
    $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{J}_{\text{REVA-PO}}(\theta; \{o_i\}, \pi_{\text{ref}})$
    if step mod $N = 0$ **then**
      $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$ ▷ synchronize old (sampling) policy
  evaluate π_θ on validation set to obtain s_{val}
  if s_{val} is the new best **then**
    $\theta_{\text{best}} \leftarrow \theta$
  if epoch mod $M = 0$ **then** ▷ Sec. 3.3
    $\pi_{\text{ref}} \leftarrow \pi_{\theta_{\text{best}}}$
    $\pi_\theta \leftarrow \pi_{\theta_{\text{best}}}$
    $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$ ▷ reset and synchronize all policies

Output: optimized policy π_θ

Figures 2c and 2d illustrate this intuition: treating π_{ref} as an anchor, π_θ as a movable ball, and $\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$ as an elastic string. A fixed anchor tightens the string as the ball moves away, confining exploration locally. By contrast, VAPR relocates the anchor to the best-performing region discovered so far. The ball then explores around this updated position, effectively extending the reachable optimization path without straying into unstable regions.

4 Experiments

4.1 Datasets and Metrics

We evaluate on two widely used CXR-RG benchmarks: MIMIC-CXR [21] and IU-Xray [10]. Following prior work [26, 40, 45], we report natural language generation (NLG) and clinical efficacy (CE) metrics.

MIMIC-CXR. The MIMIC-CXR dataset, collected at Beth Israel Deaconess Medical Center, contains 377,110 chest radiographs and 227,827 radiology reports. We use the official train/validation/test split for fair comparison with prior work.

IU-Xray. The IU-Xray dataset, developed by Indiana University, contains 7,470 chest radiographs and 3,955 reports. Following prior work [7, 24, 48], we use the established 7:1:2 train/validation/test split.

Metrics. We assess the descriptive accuracy of generated reports with NLG metrics: BLEU [31], ROUGE-L [25], and METEOR [3]. For CE metrics, we

apply the CheXpert[1] labeler to extract 14 disease categories from generated and reference reports. We binarize labels (positive as 1; negative, uncertain, and not mentioned as 0) and compute precision, recall, and F1. We also use RadGraph F1 [9] to evaluate the consistency of clinical entities (e.g., anatomical structures and lesion types) and their relationships in the generated reports.

4.2 Implementation Details

In the preliminary study, we compared two popular open-source LVLMs: Qwen2.5-VL-7B-Instruct [1] and Llama-3.2-11B-Vision-Instruct [11]. While both models achieved similar performance in Stage 1 and Stage 2, Llama required significantly more memory and was only stable using bfloat16 (BF16); float16 (FP16) training resulted in exploding gradients. In contrast, Qwen consumed less memory and maintained stability with both BF16 and FP16. Consequently, we adopted Qwen2.5-VL-7B-Instruct as our base model. All experiments were conducted on 8 NVIDIA Tesla V100 GPUs using FP16. The disease classifier, fine-tuned from [52], remained frozen during Stage 2 and Stage 3. Full implementation details and hyperparameters are available in the Supplementary Material.

4.3 Quantitative Results

NLG Metrics. Table 1 compares REVA-PO against representative and state-of-the-art methods on the MIMIC-CXR and IU-Xray datasets. Details of these methods are provided in Supplementary Material H. REVA-PO ranks first in five of six NLG metrics on IU-Xray and four of six on MIMIC-CXR. Specifically, on IU-Xray, our method surpasses strong baselines such as CMCRL [5] and AM-MRG [46] across BLEU-1 through BLEU-4 and ROUGE-L. Notably, BLEU-3 and BLEU-4 increase by +3.2% (0.253 $\rightarrow$ 0.261) and +3.6% (0.192 $\rightarrow$ 0.199), respectively. We observe similar trends on MIMIC-CXR, where REVA-PO outperforms the previous best, AM-MRG [46], with relative gains of +4.3% for BLEU-3 (0.187 $\rightarrow$ 0.195) and +5.1% for BLEU-4 (0.136 $\rightarrow$ 0.143).

Among methods evaluated on both datasets, most LLM-based approaches (MambaXray-VL [45], STREAM [58], CoD [18], B-LLM [26]) exhibit strong performance on the large-scale MIMIC-CXR but fail to surpass the non-LLM method CMCRL [5] on the smaller IU-Xray. This performance drop likely stems from a susceptibility to overfitting when training specialized image encoders on limited medical data. In contrast, our approach leverages an LVLM without the need for additional image-encoder pre-training, thereby mitigating overfitting risks. REVA-PO consistently outperforms both LLM and non-LLM baselines across both datasets, demonstrating superior generalization across varying data scales and complexities.

[1] https://github.com/stanfordmlgroup/chexpert-labeler

Table 1: Comparisons of NLG metrics between our REVA-PO and state-of-the-art methods on IU-Xray (top) and MIMIC-CXR (bottom) using BLEU-1 to BLEU-4 (B-1–4), ROUGE-L (R-L), and METEOR (MTR). Best scores are in bold.

Datasets	Types	Methods	Years	B-1	B-2	B-3	B-4	R-L	MTR
IU	non-LLM	Clinical-BERT [54]	2022	0.495	0.330	0.231	0.170	0.376	0.209
		CMMRL [32]	2022	0.494	0.321	0.235	0.181	0.384	0.201
		DCL [24]	2023	–	–	–	0.163	0.383	0.193
		METransformer [48]	2023	0.483	0.322	0.228	0.172	0.380	0.192
		PromptMRG [19]	2024	0.401	–	–	0.098	0.281	0.160
		MA [40]	2024	0.501	0.328	0.230	0.170	0.386	0.213
		$CMCRL_3$ [5]	2025	0.505	0.334	0.245	0.190	0.394	0.210
	LLM	B-LLM [26]	2024	0.499	0.323	0.238	0.184	0.390	0.208
		CoD [18]	2025	0.403	–	–	0.091	0.288	–
		STREAM [58]	2025	0.499	0.333	0.238	0.178	0.377	0.213
		MambaXray-VL [45]	2025	0.491	0.330	0.241	0.185	0.371	0.216
		AM-MRG [46]	2026	0.489	0.339	0.253	0.192	0.384	**0.225**
	LVLM	REVA-PO (ours)		**0.517**	**0.356**	**0.261**	**0.199**	**0.406**	0.216
MIMIC	non-LLM	Clinical-BERT [54]	2022	0.383	0.230	0.151	0.106	0.275	0.144
		CMMRL [32]	2022	0.381	0.232	0.155	0.109	0.287	0.151
		DCL [24]	2023	–	–	–	0.109	0.284	0.150
		METransformer [48]	2023	0.386	0.250	0.169	0.124	0.291	0.152
		NADM [64]	2023	0.402	0.258	0.179	0.130	0.289	0.155
		PromptMRG [19]	2024	0.398	–	–	0.112	0.268	0.157
		MA [40]	2024	0.396	0.244	0.162	0.115	0.274	0.151
		RRG-DPO [29]	2025	–	–	–	0.112	0.286	0.166
		$CMCRL_6$ [5]	2025	0.400	0.245	0.165	0.119	0.280	0.150
	LLM	RGRG [42]	2023	0.373	0.249	0.175	0.126	0.264	0.168
		HERGen [43]	2024	0.395	0.248	0.169	0.122	0.285	0.156
		B-LLM [26]	2024	0.402	0.262	0.180	0.128	0.291	**0.175**
		CoD [18]	2025	0.412	–	–	0.129	0.286	–
		STREAM [58]	2025	0.420	0.267	0.184	0.133	0.291	0.164
		MambaXray-VL [45]	2025	0.422	0.268	0.184	0.133	0.289	0.167
		AM-MRG [46]	2026	**0.426**	0.271	0.187	0.136	0.291	0.174
	LVLM	REVA-PO (ours)		0.424	**0.280**	**0.195**	**0.143**	**0.293**	0.170

CE Metrics. Table 2 presents the CE metrics on MIMIC-CXR, where REVA-PO establishes new benchmarks in Chexpert precision (0.569), Chexpert F1 (0.506), and RadGraph F1 (0.246). Regarding diagnostic accuracy, our method surpasses AM-MRG [46] by 4.5% in CheXpert F1 (0.484 $\rightarrow$ 0.506). Moreover, REVA-PO outperforms the previous state-of-the-art, STREAM [58], by 12.8% in RadGraph F1 (0.218 $\rightarrow$ 0.246). This substantial margin underscores that our approach generates reports with superior clinical factual consistency with reference reports.

4.4 Ablation Study

We conduct ablation studies to quantify the individual contributions of our three-stage pipeline and the modifications introduced in the RL stage (RER and VAPR). Additional stability analyses are available in the Supplementary Material.

Table 2: Comparisons of the CE metrics on MIMIC-CXR using CheXpert precision (P), CheXpert recall (R), CheXpert F1 (F1), and RadGraph F1 (RG). Best scores are in bold.

Methods	P	R	F1	RG
CMMRL [32]	0.342	0.294	0.292	–
METransformer [48]	0.346	0.309	0.311	–
MambaXray-VL [45]	0.371	0.321	0.340	–
HERGen [43]	0.415	0.301	0.371	–
DCL [24]	0.471	0.352	0.373	–
MA [40]	0.411	0.398	0.389	–
$CMCRL_6$ [5]	0.489	0.340	0.401	–
Clinical-BERT [54]	0.387	0.435	0.415	–
NADM [64]	0.417	0.413	0.415	0.209
RRG-DPO [29]	0.462	0.487	0.443	0.207
RGRG [42]	0.461	0.475	0.447	–
STREAM [58]	0.531	0.407	0.460	0.218
B-LLM [26]	0.465	0.482	0.473	–
PromptMRG [19]	0.501	0.509	0.476	–
CoD [18]	0.487	**0.521**	0.479	–
AM-MRG [46]	0.555	0.429	0.484	–
REVA-PO (ours)	**0.569**	0.455	**0.506**	**0.246**

Effect of the Three-Stage Pipeline. From Tab. 3, Stage 2 improves CE metrics by 13.5% compared to Stage 1, a gain primarily attributable to the integration of disease cues from the classifier. Notably, recall rises from 0.353 to 0.431 (+22.1%), suggesting that the classifier effectively mitigates the issue of missed findings.

Stage 3 (RL) further enhances performance, yielding a 15.6% increase in NLG metrics and a 22.1% improvement in CE metrics over the Stage 1 baseline. These results indicate that RL can simultaneously improve linguistic overlap with reference reports and diagnostic accuracy.

The final three rows of Tab. 3 confirm that both Stage 1 and Stage 2 are essential for reaching peak performance.

Effect of Response-Weighted Regularization. We monitor the average reward of responses sampled from $\pi_{\theta_{\text{old}}}$ during training. As illustrated in Fig. 3a, the reward curve for RER ($\beta = \beta_i$) consistently remains above the constant-KL baseline (CKL; $\beta = \beta_{\text{base}}$). This gap suggests that response-level adaptive scaling improves sample efficiency, leading to faster convergence and a higher reward plateau. These gains are especially important in data-constrained settings such as medical report generation. This observation is further corroborated by the empirical results in Tab. 4, where the configuration using RER (row 1) achieves higher final scores than the CKL counterpart (row 3).

Although some recent studies advocate removing the KL regularizer to encourage exploration [8, 15, 55, 59], we observe training collapse in its absence. Figure 3b shows that removing the KL term drives sampling rewards toward zero early in training, likely due to the policy drifting into low-density regions of the LVLM's prior. This confirms that KL regularization remains indispensable for stable RL in the CXR-RG task.

Table 3: Ablation study of three-stage training pipeline on MIMIC-CXR. NLG Metrics: BLEU-1 to BLEU-4 (B-1–4), ROUGE-L (R-L), and METEOR (MTR). CE Metrics: CheXpert precision (P), CheXpert recall (R), CheXpert F1 (F1), and RadGraph F1 (RG). Δ_{NLG} and Δ_{CE} denote gains over Stage 1 (S1) for NLG and CE metrics.

S1	S2	S3	B-1	B-2	B-3	B-4	R-L	MTR	Δ_{NLG}	P	R	F1	RG	Δ_{CE}
✓			0.383	0.236	0.157	0.109	0.268	0.149	–	0.480	0.353	0.407	0.215	–
✓	✓		0.401	0.251	0.167	0.118	0.277	0.156	+5.2%	0.526	0.431	0.474	0.220	+13.5%
✓		✓	0.407	0.261	0.179	0.128	0.288	0.157	+9.1%	0.516	0.393	0.446	0.228	+8.8%
	✓	✓	0.415	0.271	0.185	0.131	0.289	0.164	+11.8%	0.496	0.467	0.481	0.242	+15.9%
✓	✓	✓	0.424	0.280	0.195	0.143	0.293	0.170	+15.6%	0.569	0.455	0.506	0.246	+22.1%

Table 4: Ablation study of RL-stage modifications on MIMIC-CXR. NLG Metrics: BLEU-1 to BLEU-4 (B-1–4), ROUGE-L (R-L), and METEOR (MTR). CE Metrics: CheXpert precision (P), CheXpert recall (R), CheXpert F1 (F1), and RadGraph F1 (RG). Abbreviations: RER = Response-Weighted Regularization; CKL = constant KL; VAPR = Validation-Anchored Policy Reset. Best scores are in bold.

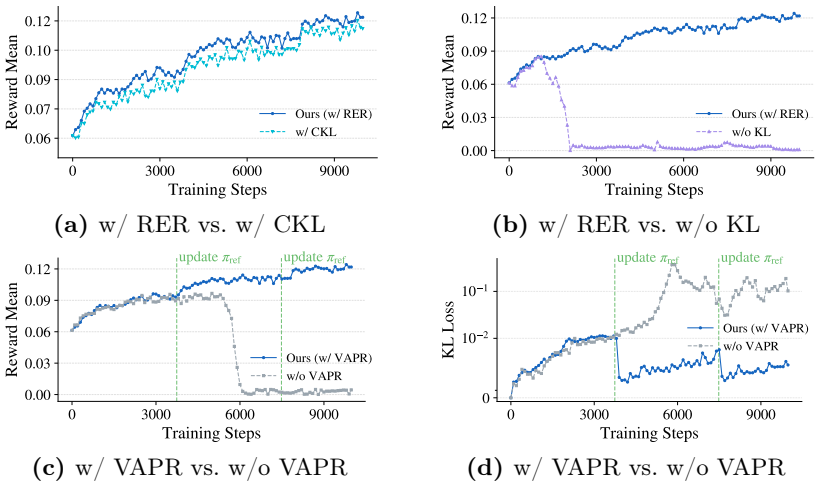
Modifications	B-1	B-2	B-3	B-4	R-L	MTR	P	R	F1	RG
Ours (RER+VAPR)	**0.424**	**0.280**	**0.195**	**0.143**	0.293	**0.170**	**0.569**	**0.455**	**0.506**	**0.246**
w/ CKL, w/o RER	0.419	0.276	0.192	0.138	**0.294**	0.168	0.528	0.448	0.484	0.239
w/o KL	0.414	0.267	0.184	0.132	0.290	0.165	0.522	0.416	0.463	0.235
w/o VAPR	0.417	0.272	0.187	0.134	0.289	0.167	0.530	0.439	0.481	0.241

Effect of Validation-Anchored Policy Reset. As seen in Fig. 3c, omitting VAPR leads to training collapse at approximately 6,000 steps. Upon the first reference update, a clear performance bifurcation occurs: the baseline without VAPR plateaus and subsequently degrades, whereas the VAPR-enabled policy continues its upward trajectory. Figure 3d further reveals that each policy reset effectively extinguishes accumulated regularization pressure, preventing the KL term from dominating the gradients. Consistently, Tab. 4 confirms that VAPR yields higher scores across all metrics (row 1 *vs*. row 4), validating its role in sustaining long-term optimization.

4.5 Qualitative Results

Figure 4 compares outputs across the three stages and the baseline (Qwen2.5-VL-7B-Instruct without fine-tuning). Adding the classifier in Stage 2 helps identify clinically relevant cues that the baseline and Stage 1 miss. RL in Stage 3 improves wording precision and reduces hallucinated content.

Case 1. The baseline and Stage 1 miss lung abnormalities. With disease classifier guidance, Stage 2 identifies *lobar consolidation* but assigns it to the wrong location (*left lower*) and uses vague phrasing. Stage 3 corrects these issues: it describes the finding as *airspace opacity*, localizes it to the *right lower lung*, and infers *pneumonia*, consistent with the reference.



(a) w/ RER vs. w/ CKL **(b)** w/ RER vs. w/o KL

(c) w/ VAPR vs. w/o VAPR **(d)** w/ VAPR vs. w/o VAPR

Fig. 3: Ablation study of RL-stage modifications on MIMIC-CXR. Mean reward of responses sampled from $\pi_{\theta_{\text{old}}}$ *vs.* training steps. (a) Response-Weighted Regularization (RER) outperforms constant KL (CKL). (b) Omitting the KL term causes early collapse. (c) Validation-Anchored Policy Reset (VAPR) prevents collapse and boosts rewards; green lines mark π_{ref} updates. (d) KL loss *vs.* training steps, plotted on a symmetric logarithmic scale to accommodate the wide value range in later training.

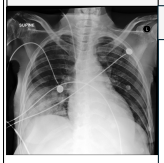
Input Image	Reference Report	Qwen2.5-VL-7B-Instruct	Stage 1	Stage 2	Stage 3
	**GT Categories:** Consolidation, Pneumonia, Support Devices Right ij central venous catheter is seen with its tip in the expected location of the mid svc. ① There is airspace consolidation in the right lower lung concerning for pneumonia. ② The left lung is mostly clear. No large effusion is seen. No pneumothorax. Cardiomediastinal silhouette is stable. Bony structures are intact.	**Pred Categories:** N/A ... 2. Lungs: The lung fields appear clear without evidence of consolidation, masses, or pleural effusions. There is no significant evidence of atelectasis or pneumothorax. ② 3. Pleura: The costophrenic angles are sharp, indicating no pleural effusion. 4. Heart: The heart size appears within normal limits for this view... 8. Bones: There are no obvious fractures or lytic/sclerotic lesions...	**Pred Categories:** N/A The lungs are clear. The cardiomediastinal silhouette is within normal limits. No acute osseous abnormalities identified. No free air below the right hemidiaphragm. No pneumothorax. No acute osseous abnormalities identified. No free air below the right hemidiaphragm. No pneumothorax. No free air below the right hemidiaphragm. No pneumothorax.	**Pred Categories:** Consolidation, Pneumonia, Lung Opacity, Support Devices A right internal jugular central venous catheter terminates in the mid svc. ① There is no pneumothorax. The cardiomediastinal silhouette is within normal limits. There is a new left lower lobe consolidation. ② There is no large pleural effusion. No acute osseous abnormalities identified. There is no evidence of pneumoperitoneum. A right pice terminates in the low svc.	**Pred Categories:** Consolidation, Pneumonia, Lung Opacity, Support Devices There is a right internal jugular central venous catheter with tip in the mid svc. ① There is no pneumothorax. There is a new airspace opacity in the right lower lung concerning for pneumonia. ② The left lung is clear. There is no pleural effusion. The cardiac silhouette is normal. The mediastinal contours are normal. There is no free air under the right hemidiaphragm.
	**GT Categories:** Cardiomegaly, Support Devices Right picc terminates in mid svc. Left pectoral pacemaker has its leads terminating in right atrium and right ventricle. ① Cardiac silhouette is mildly enlarged. ② Prosthetic heart valve and median sternotomy wires are in unchanged position. There is no consolidation pleural effusion or pneumothorax.	**Pred Categories:** N/A ... 3. Pleura: There are no visible pleural effusions or thickening. 4. Heart: The heart size appears within normal limits. The cardiac silhouette is well-defined, and there are no signs of cardiomegaly or pericardial effusion. ② ... 9. Support Devices: There is a pacemaker visible in the right chest region, which is a support device. ① ...	**Pred Categories:** N/A The patient is status post median sternotomy and cabg. Left-sided dual-chamber pacemaker device is noted with leads terminating in the right atrium and right ventricle unchanged. ① Moderate enlargement of the cardiac silhouette is re-demonstrated. ② The aorta is mildly tortuous and diffusely calcified. No focal consolidation pleural effusion or pneumothorax is present. There are mild bibasilar atelectasis. No acute osseous abnormalities are identified.	**Pred Categories:** Cardiomegaly, Support Devices The patient is status post median sternotomy and cabg. Left-sided pacemaker device is noted with leads terminating in the right atrium and right ventricle. ① Moderate enlargement of cardiac silhouette is re-demonstrated. ② The aorta remains tortuous. There is mild pulmonary vascular congestion without overt pulmonary edema. No focal consolidation pleural effusion or pneumothorax is identified. No acute osseous abnormalities seen.	**Pred Categories:** Cardiomegaly, Support Devices Frontal and lateral views of the chest were obtained. There is no focal consolidation pleural effusion or pneumothorax. The cardiac silhouette is mildly enlarged. ② The mediastinal contours are unremarkable. The patient is status post median sternotomy with intact wires. A left pectoral pacemaker is in unchanged position with leads terminating in the right atrium and right ventricle. ①

Fig. 4: Qualitative comparison examples on MIMIC-CXR: baseline (Qwen2.5-VL-7B-Instruct) and Stage 1 to Stage 3. Columns: input image; reference with ground-truth (GT) categories; and outputs with classifier-predicted categories. Sentences referring to the same anatomical region are color-coded and key phrases are underlined.

Case 2. The baseline detects the *pacemaker* but misses heart enlargement. Stage 1 and Stage 2 overestimate cardiac size as *moderate* enlargement and introduce several hallucinations, including *tortuous aorta*, *atelectasis*, and *mild pulmonary vascular congestion*. Stage 3 removes all hallucinations and correctly describes the cardiac silhouette as *mildly* enlarged.

5 Conclusion

We present REVA-PO, a policy optimization framework for LVLM-based CXR report generation. By integrating Response-Weighted Regularization (RER) with Validation-Anchored Policy Reset (VAPR), REVA-PO enhances training stability and exploratory capacity. RER dynamically scales response-level penalties via advantages and reference entropy, while VAPR periodically resets policies to the best-performing anchor to alleviate accumulated regularization pressure. Evaluations on MIMIC-CXR and IU-Xray confirm that REVA-PO achieves state-of-the-art performance in both linguistic fluency and clinical accuracy.

Acknowledgements

We thank Jianfeng Wang from the Department of Radiology at The First People's Hospital of Zhengzhou for his assistance with the radiologist evaluation in this study. We also acknowledge UBC Advanced Research Computing and the Digital Research Alliance of Canada for providing computational resources that supported this work.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 2, 4, 10
2. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022) 2, 4
3. Banerjee, S., Lavie, A.: 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. pp. 65–72 (2005) 9
4. Berlin, L.: Liability of interpreting too many radiographs. American Journal of Roentgenology **175**(1), 17–22 (2000) 1
5. Chen, W., Liu, Y., Wang, C., Zhu, J., Li, G., Liu, C.L., Lin, L.: Cross-modal causal representation learning for radiology report generation. IEEE Transactions on Image Processing (2025) 4, 10, 11, 12
6. Chen, Y., Ge, Y., Wang, R., Ge, Y., Cheng, J., Shan, Y., Liu, X.: Grpo-care: Consistency-aware reinforcement learning for multimodal reasoning. arXiv preprint arXiv:2506.16141 (2025) 2, 4
7. Chen, Z., Shen, Y., Song, Y., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Aug 2021) 9
8. Cui, G., Yuan, L., Wang, Z., Wang, H., Li, W., He, B., Fan, Y., Yu, T., Xu, Q., Chen, W., et al.: Process reinforcement through implicit rewards. arXiv preprint arXiv:2502.01456 (2025) 12

9. Delbrouck, J.B., Chambon, P., Chen, Z., Varma, M., Johnston, A., Blankemeier, L., Van Veen, D., Bui, T., Truong, S., Langlotz, C.: Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 12902–12915 (2024) 8, 10
10. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2015) 9
11. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) 2, 10
12. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**(8081), 633–638 (2025) 2, 4
13. Guo, L., Tahir, A.M., Zhang, D., Wang, Z.J., Ward, R.K., et al.: Automatic medical report generation: Methods and applications. APSIPA Transactions on Signal and Information Processing **13**(1) (2024) 1
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022) 6
15. Hu, J., Zhang, Y., Han, Q., Jiang, D., Zhang, X., Shum, H.Y.: Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. arXiv preprint arXiv:2503.24290 (2025) 12
16. Huang, Z., Zhang, X., Zhang, S.: Kiut: Knowledge-injected u-transformer for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19809–19818 (2023) 4
17. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019) 6
18. Jin, H., Che, H., He, S., Chen, H.: A chain of diagnosis framework for accurate and explainable radiology report generation. IEEE Transactions on Medical Imaging (2025) 1, 4, 10, 11, 12
19. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2607–2615 (2024) 3, 6, 11, 12
20. Jing, B., Wang, Z., Xing, E.: Show, describe and conclude: On exploiting the structure information of chest X-ray reports. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6570–6580 (2019) 2
21. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019) 9
22. Krupinski, E.A., Berbaum, K.S., Caldwell, R.T., Schartz, K.M., Kim, J.: Long radiology workdays reduce detection and accommodation accuracy. Journal of the American College of Radiology **7**(9), 698–704 (2010) 1
23. Lee, S., Han, J., Song, H., Choi, S.J., Lee, H., Yu, Y.: KL penalty control via perturbation for direct preference optimization. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025) 2, 4

24. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3334–3343 (2023) 4, 9, 11, 12
25. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004) 9
26. Liu, C., Tian, Y., Chen, W., Song, Y., Zhang, Y.: Bootstrapping large language models for radiology report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 18635–18643 (2024) 1, 4, 9, 10, 11, 12
27. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2021) 4
28. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024) 5
29. Liu, H., Wei, D., Xu, Z., Wu, X., Zheng, Y., Wang, L.: Rrg-dpo: Direct preference optimization for clinically accurate radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 552–562. Springer (2025) 11, 12
30. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022) 2, 4
31. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002) 9
32. Qin, H., Song, Y.: Reinforced cross-modal alignment for radiology report generation. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 448–458 (2022) 2, 4, 11, 12
33. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023) 4
34. Ramé, A., Ferret, J., Vieillard, N., Dadashi, R., Hussenot, L., Cedoz, P.L., Sessa, P.G., Girgin, S., Douillard, A., Bachem, O.: Warp: On the benefits of weight averaged rewarded policies. arXiv preprint arXiv:2406.16768 (2024) 2, 4
35. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7008–7024 (2017) 4
36. Rimmer, A.: Radiologist shortage leaves patient care at risk, warns royal college. BMJ: British Medical Journal (Online) **359** (2017) 1
37. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017) 2, 4
38. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024) 2, 4
39. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025) 4
40. Shen, H., Pei, M., Liu, J., Tian, Z.: Automatic radiology reports generation via memory alignment network. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4776–4783 (2024) 9, 11, 12

41. Sloan, P., Clatworthy, P., Simpson, E., Mirmehdi, M.: Automated radiology report generation: A review of recent advances. IEEE Reviews in Biomedical Engineering **18**, 368–387 (2024) 1
42. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7433–7442 (2023) 11, 12
43. Wang, F., Du, S., Yu, L.: Hergen: Elevating radiology report generation with longitudinal data. In: European Conference on Computer Vision. pp. 183–200. Springer (2024) 1, 4, 11, 12
44. Wang, L., Ning, M., Lu, D., Wei, D., Zheng, Y., Chen, J.: An inclusive task-aware framework for radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 568–577. Springer (2022) 3
45. Wang, X., Wang, F., Li, Y., Ma, Q., Wang, S., Jiang, B., Tang, J.: Cxpmrg-bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5123–5133 (2025) 1, 4, 9, 10, 11, 12
46. Wang, X., Wang, F., Wang, H., Jiang, B., Li, C., Wang, Y., Tian, Y., Tang, J.: Activating associative disease-aware vision token memory for llm-based x-ray report generation. IEEE Transactions on Medical Imaging **45**, 583–595 (2026) 4, 10, 11, 12
47. Wang, Z., Han, H., Wang, L., Li, X., Zhou, L.: Automated radiographic report generation purely on transformer: A multicriteria supervised approach. IEEE Transactions on Medical Imaging **41**(10), 2803–2813 (2022) 3
48. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11558–11567 (2023) 9, 11, 12
49. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 655–664. Springer (2022) 3, 4
50. Wang, Z., Zhou, L., Wang, L., Li, X.: A self-boosting framework for automated radiographic report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2433–2442 (2021) 4
51. Wu, J., Xie, Y., Yang, Z., Wu, J., Gao, J., Ding, B., Wang, X., He, X.: β-dpo: Direct preference optimization with dynamic β. Advances in Neural Information Processing Systems pp. 129944–129966 (2024) 2, 4
52. Xiao, J., Bai, Y., Yuille, A., Zhou, Z.: Delving into masked autoencoders for multi-label thorax disease classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3588–3600 (2023) 10
53. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2087–2098 (2025) 2
54. Yan, B., Pei, M.: Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 2982–2990 (2022) 11, 12
55. Yan, J., Li, Y., Hu, Z., Wang, Z., Cui, G., Qu, X., Cheng, Y., Zhang, Y.: Learning to reason under off-policy guidance. arXiv preprint arXiv:2504.14945 (2025) 12

56. Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S.K., Xiao, L.: Radiology report generation with a learned knowledge base and multi-modal alignment. Medical Image Analysis **86**, 102798 (2023) 3
57. Yang, S., Wu, X., Ge, S., Zhou, S.K., Xiao, L.: Knowledge matters: Chest radiology report generation with general and specific knowledge. Medical image analysis **80**, 102510 (2022) 4
58. Yang, Y., You, X., Zhang, K., Fu, Z., Wang, X., Ding, J., Sun, J., Yu, Z., Huang, Q., Han, W., et al.: Spatio-temporal and retrieval-augmented modelling for chest x-ray report generation. IEEE Transactions on Medical Imaging (2025) 1, 4, 10, 11, 12
59. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476 (2025) 7, 12
60. Zang, Y., Dong, X., Zhang, P., Cao, Y., Liu, Z., Ding, S., Wu, S., Ma, Y., Duan, H., Zhang, W., et al.: Internlm-xcomposer2.5-reward: A simple yet effective multi-modal reward model. arXiv preprint arXiv:2501.12368 (2025) 4
61. Zhan, Y., Zhu, Y., Zheng, S., Zhao, H., Yang, F., Tang, M., Wang, J.: Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. arXiv preprint arXiv:2503.18013 (2025) 4
62. Zhang, Y., Yu, T., Tian, H., Fu, C., Li, P., Zeng, J., Xie, W., Shi, Y., Zhang, H., Wu, J., Wang, X., Hu, Y., Wen, B., Gao, T., Zhang, Z., Yang, F., ZHANG, D., Wang, L., Jin, R.: MM-RLHF: The next step forward in multimodal LLM alignment. In: Forty-second International Conference on Machine Learning (2025) 4
63. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12910–12917 (2020) 4
64. Zhao, G., Yan, Y., Zhao, Z.: Normal-abnormal decoupling memory for medical report generation. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 1962–1977 (2023) 11, 12
65. Zheng, R., Shen, W., Hua, Y., Lai, W., Dou, S., Zhou, Y., Xi, Z., Wang, X., Huang, H., Gui, T., Zhang, Q., Huang, X.: Improving generalization of alignment with human preferences through group invariant learning. In: International Conference on Learning Representations. pp. 44714–44736 (2024) 2, 4
66. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) 5
67. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019) 2, 4