RHDP: A GRADIENT-VARIATION-AWARE REWARD HACKING DETECTION AND PENALIZATION MECHANISM FOR CHEST X-RAY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Chest X-ray report generation aims to produce a detailed and clinically accurate description for a given image. Although reinforcement learning (RL) fine-tuning can optimize sequence-level preferences, an imperfect proxy reward may be over-optimized, leading to template collapse and deterioration of the clinically meaningful gold objective. We introduce RHDP, a reward-hacking detection and penalization framework built around the long-window variance of input gradients (LVoG). LVoG is the temporal variance of the absolute input gradient of the joint CE–RL loss and is used as a sample-wise reliability score: abnormally low values indicate increased proxy–gold misalignment risk. A decomposition analysis shows that LVoG combines CE–RL transition dynamics, reward-side advantage dynamics, and pure RL policy-side image-conditioned sensitivity; the policy-side signal remains predictive after removing instantaneous advantage scaling. RHDP uses lagged LVoG to modulate the reward and the forward KL divergence, suppressing unreliable updates while preserving generation quality. Experiments on IU-Xray and MIMIC-CXR demonstrate improved clinical accuracy, competitive semantic quality, and consistent mitigation of reward hacking. The code is available on GitHub.

1 INTRODUCTION

Automatic chest X-ray report generation (RRG) is an important clinical task that can reduce the workload of radiologists in reading images and writing reports (Vera et al., 2020; Zhang et al., 2023c). However, the complex physiological structures and pathological features in chest X-rays are difficult to align with medical report features, making this task challenging. Most researchers optimize encoder-decoder architectures with the cross-entropy (CE) loss (Tanida et al., 2023; Yan et al., 2022), which measures only local word-level differences and can hardly induce the unique clinical semantics of medical texts. To mitigate this, another line of work adopts self-critical sequence training (SCST) (Rennie et al., 2017) with a verifiable reward (Qin & Song, 2022; Jing et al., 2019). However, proxy reward functions cannot perfectly characterize the real objective because of their loopholes (Skalse et al., 2022); through backpropagation, the model may manipulate the reward, narrowing the output distribution and collapsing into a certain mode (Krishnan & Evans, 2025; Zhang et al., 2025), usually concealed. Inspired by Skalse et al. (2022); Laidlaw et al. (2025), we use the simple prior that the proxy reward and the gold reward should be optimized consistently during training to measure the hacking behavior of the model. As shown in Fig. 1, the current mainstream RRG fine-tuning methods (Qin & Song, 2022; Yi et al., 2025) all over-optimize the proxy reward and degrade the gold reward. Although Ostmeier et al. (2024); Yu et al. (2023) design more clinically applicable reward functions to mitigate reward over-optimization, they do not systematically study the mechanism of reward hacking in RRG models.

To solve this problem, it is necessary to capture and quantify the reward hacking phenomenon during RL training. Models often exploit shortcuts to achieve high rewards, which manifests as distinctive phenomena at the model feature level (Miao et al., 2025; Wang et al., 2026). Based on this observation, we propose a gradient-variation-aware hacking detector. Inspired by saliency maps (Simonyan et al., 2014) and energy loss (Miao et al., 2025), we compute the absolute gradient of the joint CE–RL loss with respect to the input image and track the variance of its spatial mean over a long

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

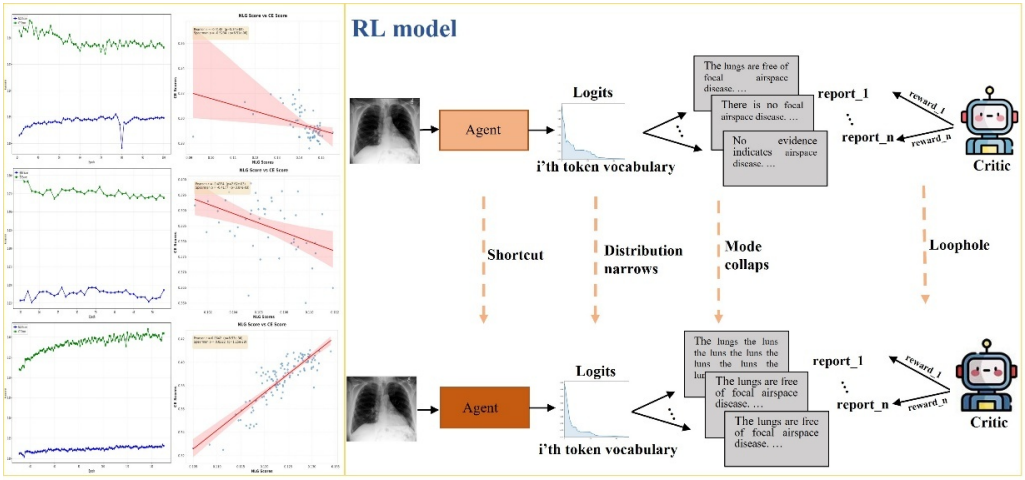

Figure 1: The left panel shows that mainstream RL methods (ViT-RL, LHR-RFL) exhibit a negative correlation between proxy reward (text overlap) and gold reward (clinical accuracy), while our RHDP constraint effectively restores this to a positive correlation. For a fair comparison, we keep the reward function of these mainstream models (i.e., the weighted NLG reward) unchanged and apply the proposed RHDP on top; although the alignment improvement attainable under this reward is limited, RHDP still mitigates the inconsistency of its optimization. The right panel is a conceptual schematic illustrating the resulting output distribution collapse, characterized by length over-optimization and repetitive generation.

causal window. The resulting statistic, LVoG (long-window variance of gradients), is a sample-wise gradient-dynamics reliability indicator: within the windowing regime used in our experiments, larger LVoG is associated with stronger proxy–gold optimization consistency, whereas abnormally low LVoG indicates higher reward-misalignment risk. Based on the quantified reliability signal, we reweight the reward in a focal-like form (Lin et al., 2017) to suppress hacking behavior. Meanwhile, we modulate the forward KL divergence to consolidate genuinely well-aligned samples while selectively relaxing (or zeroing) it on hacking samples to preserve exploration. By analyzing the possible loopholes of the mainstream reward functions used in existing studies, including weighted NLG (Yi et al., 2025), BERTScore (Zhang et al., 2020), RadGraph (Miura et al., 2021), and RadCliQ (Yu et al., 2023), we evaluate reward hacking from five complementary aspects (detailed in Section 4.1).

The contributions of this work are summarized as follows.

- We propose a gradient-variation-aware reward hacking detector that captures hacking behavior from the model's overall internal dynamics.
- Based on the captured hacking phenomenon and the optimization state of samples, we adopt a unified reward modulation strategy and a divergence modulation strategy to suppress the sampling behavior of hacking samples while consolidating perfect samples.
- We improve the RadCliQ reward and evaluate the hacking behavior of end-to-end RRG models trained with SCST under mainstream verifiable rewards, using whole-dataset statistics rather than individual samples. Experiments on two benchmarks verify leading semantic fluency, clinical accuracy, and hacking mitigation.

2 Related Work

2.1 Different Optimization Strategies for Encoder-Decoder Architectures

Inspired by image captioning (Zhang et al., 2023b;a), researchers use CNN-RNN or CNN-Transformer encoder-decoder architectures, mostly optimized with the CE loss. For example, Chen et al. (2020) use memory matrix modeling for long-range expression, and Tang et al. (2025) use multi-scale feature modeling for multi-granularity alignment. However, CE-optimized models lack

clinical preference. The SCST paradigm, which directly optimizes text-fluency or clinical-accuracy metrics, improves performance (Rennie et al., 2017; Qin & Song, 2022) but often over-optimizes the reward, causing the model to ignore the input and collapse into a certain distribution.

2.2 Reward Hacking in Reinforcement Learning

RL fine-tuning based on verifiable rewards (Shao et al., 2024; Yu et al., 2025) is an important stage for large language models, but reward functions often contain exploitable loopholes. Miao et al. (2025) quantify reward hacking by the excessive energy-loss increase of the last layer and constrain the reward accordingly. Laidlaw et al. (2025) quantify hacking by the proxy–gold reward correlation and mitigate it with occupancy-measure regularization. Wang et al. (2026) capture hacking from neuron activations via gradient fingerprints, and Li et al. (2025) replace reverse KL with forward KL to consolidate high-quality LLM samples. In RRG, Yu et al. (2023) emphasize clinically aligned reward design, and Wang et al. (2025) link reward exploitation to length anomalies. These methods construct reward functions from different perspectives but do not directly address reward hacking or the mode collapse caused by reward loopholes.

3 Method

3.1 Overview

Figure 2 presents the RHDP framework. Built on an encoder-decoder backbone with RL, we follow (Chen et al., 2020; Qin & Song, 2022) to flatten visual grid features into a sequence and auto-regressively decode the report $Y = \mathcal{D}(\mathcal{E}(f_{\text{ve}}(I)))$, where f_{ve} is a pre-trained ResNet101 (He et al., 2016), and $\mathcal{E}$ and $\mathcal{D}$ are the encoder and decoder. The cross-entropy loss is

$$\mathcal{L}_{\text{CE}}(\theta) = -\sum_{t=1}^{L} \log p_\theta(y_t \mid y_{1:t-1}), \tag{1}$$

and following SCST (Rennie et al., 2017), the RL loss is

$$\mathcal{L}_{\text{RL}}(\theta) = -\mathbb{E}_{Y\sim\pi(\theta)}[r(Y)], \tag{2}$$

where r is the verifiable reward between the generated report and the ground-truth reference, θ denotes the model parameters, and $\pi(\theta)$ is the policy. Since the reward is non-differentiable, the policy gradient (Williams, 1992) with Monte Carlo sampling estimates

$$\nabla\mathcal{L}_{\text{RL}}(\theta) \approx -(r(Y) - b)\,\nabla_\theta \log \pi_\theta(Y), \tag{3}$$

where b is the baseline reward alleviating the high variance of random sampling. The two losses are finally weighted and combined:

$$\mathcal{L}_{\text{joint}} = \alpha_1\mathcal{L}_{\text{CE}} + \alpha_2\mathcal{L}_{\text{RL}} \tag{4}$$

Following (Qin & Song, 2022), we set the balancing weights for CE diversity and RL exploration to 0.05 and 0.95, respectively.

3.2 Gradient-Variation-Aware Hacking Detector and Reward Modulation

When the model exploits a loophole in the reward function and collapses toward a shortcut mode, its input-conditioned joint training signal changes abnormally. We therefore use the long-window variance of the input gradient of the joint loss as the core indicator. For sample n at training step t, the pixel-level absolute saliency is

$$S_{n,t,i} = \left|\frac{\partial\mathcal{L}_{\text{joint},n,t}}{\partial I_{n,t,i}}\right|, \quad i = 1, 2, \ldots, M \tag{5}$$

where M is the number of input pixels, N the number of samples, and T the number of training steps. The absolute value prevents cancellation between positive and negative pixels. Unlike Agarwal et al. (2022), who use classifier-head activations, we use an end-to-end generation gradient from the joint

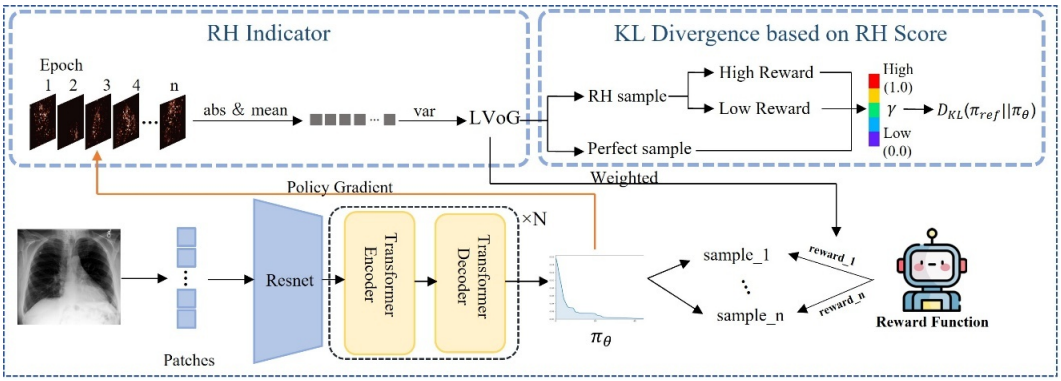

Figure 2: The overall framework of RHDP, including the gradient-variation-aware hacking indicator, the hacking-aware reward modulation and divergence modulation strategies.

objective. We then aggregate each saliency map spatially and compute its variance over a long causal window $\mathcal{W}_t$ of length K:

$$\mu_{n,t} = \frac{1}{M}\sum_{i=1}^{M} S_{n,t,i} = \frac{1}{M}\left\|\nabla_{I_n}\mathcal{L}_{\text{joint},n,t}\right\|_1 \tag{6}$$

$$V_{n,t} := \text{LVoG}_{n,t} = \text{Var}_{\tau\in\mathcal{W}_t}\left(\mu_{n,\tau}\right) \tag{7}$$

A low mean indicates weak instantaneous input responsiveness but cannot distinguish benign confidence from shortcut exploitation. The long-window variance is more informative: an abnormally low $V_{n,t}$ is associated with low proxy–gold alignment and hence high hacking risk. We therefore normalize raw LVoG as a reliability score $q_{n,t}$ and define the risk score $h_{n,t} = 1 - q_{n,t}$:

$$q_{n,t} = \frac{V_{n,t} - V_{\min}}{V_{\max} - V_{\min}} \in [0,1], \qquad h_{n,t} = 1 - q_{n,t} \tag{8}$$

A focal-like factor then controls the contribution of each sample. Lower reliability $q_{n,t}$ (equivalently, higher risk $h_{n,t}$) receives a smaller relative reward multiplier:

$$\lambda_{n,t} = (1 + q_{n,t})^{\alpha} + 1 \tag{9}$$

$$R_{\text{fin}} = \lambda \cdot R_{\text{ini}} \tag{10}$$

Here, R_{ini} and R_{fin} denote the reward values before and after adjustment, respectively.

Because the operational statistic is computed from the joint objective, it should not be written as the variance of $|A|S$ alone. The exact vector-level decomposition is

$$\nabla_x\mathcal{L}_{\text{joint}} = \alpha_1\nabla_x\mathcal{L}_{\text{CE}} + \alpha_2 A^p\nabla_x\ell^{\text{pol}}, \qquad \text{LVoG} = \text{Var}_{\tau\in\mathcal{W}_t}\left[\frac{1}{M}\left\|\nabla_x\mathcal{L}_{\text{joint},\tau}\right\|_1\right] \tag{11}$$

Thus, LVoG is a cross-objective composite indicator containing CE-side anchoring and transition dynamics, reward-side advantage dynamics, pure RL policy-side image-conditioned sensitivity, and the CE–RL directional interaction. Appendix A defines decoupled diagnostics for these channels and shows that the pure policy-side variance remains predictive of reward misalignment, whereas CE-dominated diagnostics exhibit an inverted-U relation.

3.3 Unified Divergence Modulation Strategy

When the model discovers high-reward shortcuts, the output distribution becomes sharp, leading to a loss of sampling diversity. We compute the forward KL divergence between the current model and the SFT model:

$$D_{\text{KL}}(\pi_{\text{ref}}\|\pi_\theta) = \mathbb{E}_{\pi_{\text{ref}}}\log\frac{\pi_{\text{ref}}}{\pi_\theta} \tag{12}$$

where π_{ref} is the pre-trained CE model and π_θ the current RL-fine-tuned policy. Unlike reverse KL, forward KL puts the current policy in the denominator and penalizes the loss of modes supported

by the reference distribution. We design a modulation factor γ from the LVoG reliability q and the immediate reward:

$$\gamma_{n,t} = (1 + R_{n,t})^{\beta}\ q_{n,t}^{\delta}\ \epsilon_t \tag{13}$$

Here β controls the immediate-reward influence, δ the LVoG-reliability influence, and ϵ_t is a time-dependent decay factor. When reliability is very low ($q_{n,t} \to 0$, i.e., $h_{n,t} \to 1$), the modulation approaches zero, preserving exploration instead of consolidating a suspected shortcut. When LVoG alone is not decisive, the immediate reward provides complementary information: a high-reward, high-reliability sample receives stronger consolidation (Appendix A).

3.4 Optimization Objective

We adopt the RadCliQ reward of Yu et al. (2023), a weighted combination of BERTScore (Zhang et al., 2020), CheXbert Similarity (Smit et al., 2020), and RadGraph F1 (Miura et al., 2021) that is highly consistent with physician judgments. We replace the CheXbert Similarity component with the CheXbert F1 score computed from the sample's 14-category detection results against the ground-truth report, thereby avoiding the reward loopholes introduced by the label extractor's inherent training bias. In parallel with mitigating reward hacking, we improve RadCliQ to obtain more accurate reward alignment during RL fine-tuning. Using this improved RadCliQ for the RL loss, the overall loss is:

$$\mathcal{L}_{\text{total}}(\theta) = \alpha_1 \mathcal{L}_{\text{CE}} + \alpha_2 \mathcal{L}_{\text{RL}}(\lambda R_{\text{ini}}) + \alpha_3 \mathcal{L}_{\text{KL}} \tag{14}$$

The weights are tuned as hyperparameters.

4 Experiments

4.1 Datasets and Evaluation Protocol

We evaluate on two benchmark datasets: MIMIC-CXR (473,057 images, 206,536 reports) (Johnson et al., 2019) and IU-Xray (7,470 image pairs, 3,955 reports) (Demner-Fushman et al., 2016). For efficient validation, we sample 10,000 cases from MIMIC-CXR stratified by 14 disease labels. We assess models from three perspectives. For semantic fluency, we report BLEU-4 (Papineni et al., 2002), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), METEOR (Banerjee & Lavie, 2005), and BERTScore (Zhang et al., 2020). For clinical accuracy, we use CheXbert (14-disease labeling) (Smit et al., 2020) and RadGraph (entity extraction) (Miura et al., 2021) with precision, recall, and F1 against ground truth. We evaluate reward hacking from five aspects: (1) shortcut exploitation via the disease-label distribution, measured by the precision/recall (PR) balance across 14 diseases; (2) length anomalies and repetitive generation; (3) last-layer energy loss; (4) feature-space distribution shift; and (5) output-entropy collapse.

4.2 Experimental Setup

Following previous work (Chen et al., 2020), all images are cropped to 224×224. The maximum generation lengths of the two datasets are 60 and 100, respectively, and the minimum word frequency is set to 3. The learning rates of the visual extractor and the encoder-decoder are set to 10^{-6} and 10^{-5}, respectively. The advantage function is the difference between the beam-search reward and the greedy-decoding baseline reward.

4.3 Comparison with Previous Work

Table 1 shows that our method achieves the strongest CheXbert F1 on both datasets while remaining competitive in semantic quality against CE-optimized methods (ViT (Dosovitskiy et al., 2021), R2Gen (Chen et al., 2020), PPKED (Liu et al., 2021), DCL (Li et al., 2023), ME-Transformer (Wang et al., 2023), KiUT (Huang et al., 2023), DACG (Lang et al., 2025), VSANet (Yi et al., 2026)) and RL methods (R2GenRL (Chen et al., 2020), LHR-RFL (Yi et al., 2025), CMNRL (Qin & Song, 2022), MIRGAN (Zhang & Liu, 2023)). It does not dominate every fluency metric; its main advantage is a better clinical-accuracy–fluency trade-off without extra feature-engineering modules. This is expected: penalizing shortcuts suppresses repetitive n-gram templates, lowering rigid surface-overlap scores such as BLEU-4 and CIDEr (where several CE baselines remain higher) while substantially improving semantic alignment (METEOR) and clinical fidelity (CheXbert F1). Relative to

Table 1: Performance comparison with previous work. Best results are in bold. "#" marks results taken from the original papers. Abbreviations: R-L = ROUGE-L, BERT-S = BERTScore, RadG-F1 = RadGraph F1, CheX-F1 = CheXbert F1.

Dataset	Method	BLEU-4	METEOR	R-L	CIDEr	BERT-S	RadG-F1	CheX-F1
IU-Xray	PPKED#	0.168	–	0.376	0.351	–	–	–
	DCL#	0.163	0.193	0.383	**0.586**	–	–	–
	ME-Transformer#	0.172	0.192	0.380	0.435	–	–	–
	KiUT#	0.185	0.242	0.409	–	–	–	–
	VSANet#	**0.200**	0.229	**0.425**	–	–	–	–
	MIRGAN#	0.196	0.208	0.390	0.421	–	–	–
	ViT	0.110	0.167	0.308	0.221	0.552	0.281	0.319
	R2Gen	0.148	0.178	0.345	0.420	0.550	0.291	0.334
	DACG	0.133	0.176	0.341	0.366	0.548	0.281	0.337
	CMNRL	0.137	0.175	0.332	0.277	0.566	0.314	0.347
	R2GenRL	0.182	0.218	0.358	0.361	**0.634**	**0.422**	0.497
	LHR-RFL	0.175	0.207	0.383	0.381	0.615	0.386	0.372
	Ours	0.169	**0.369**	0.383	0.401	0.614	0.395	**0.507**
MIMIC-CXR-10k	ViT	0.071	0.128	0.237	0.078	0.512	0.200	0.268
	R2Gen	0.087	0.114	0.264	0.109	0.506	0.192	0.262
	DACG	0.074	0.106	0.257	0.074	0.503	0.172	0.156
	CMNRL	0.052	0.097	0.228	0.042	0.151	0.057	0.174
	R2GenRL	**0.105**	0.124	**0.292**	**0.118**	0.518	0.226	0.148
	LHR-RFL	0.091	0.120	0.278	0.097	0.543	0.212	0.165
	Ours	0.078	**0.267**	0.252	0.105	**0.545**	**0.251**	**0.369**

Table 2: Performance of different components. Best results are in bold. Abbreviations: PR-dist = PR distance, Ent = Entropy.

Dataset	Method	BLEU-4	METEOR	R-L	CIDEr	BERT-S	RadG-F1	CheX-F1	PR-dist	Ent
IU-Xray	Base	0.161	**0.409**	0.368	0.236	0.601	0.375	0.377	0.346	0.443
	Base+LVoG	**0.182**	0.380	**0.379**	**0.349**	0.597	**0.382**	**0.461**	0.453	**0.470**
	Base+EL	**0.172**	0.193	**0.372**	**0.264**	0.597	**0.386**	**0.401**	**0.283**	**0.741**
	Base+KL	**0.171**	0.191	**0.387**	**0.432**	**0.617**	**0.430**	**0.510**	0.389	**0.594**
	Base+ALL	**0.169**	0.369	**0.383**	**0.401**	**0.614**	**0.395**	**0.507**	**0.311**	**0.503**
MIMIC-CXR-10k	Base	0.080	0.130	0.251	0.084	0.536	**0.264**	0.377	0.774	0.272
	Base+LVoG	0.062	0.123	0.230	0.080	**0.537**	0.263	**0.399**	**0.582**	**0.303**
	Base+EL	0.064	**0.152**	0.235	0.064	0.530	0.247	0.345	**0.588**	**0.280**
	Base+KL	**0.086**	**0.135**	**0.261**	**0.111**	**0.543**	0.246	0.350	**0.466**	**0.274**
	Base+ALL	0.078	**0.267**	**0.252**	**0.105**	**0.545**	0.251	0.369	**0.481**	**0.302**

SCST baselines, forward KL preserves SFT-learned modes, while LVoG-based modulation improves training stability and reduces shortcut-driven collapse. R2GenRL denotes our SCST implementation on R2Gen, with a simple weighted NLG reward.

4.4 Ablation Study

Compared with both Base and EL, our two core components jointly improve semantic fluency and clinical accuracy and maintain low hacking risk. Notably, the Base model already optimizes the improved RadCliQ reward; the additional gains from LVoG and forward KL therefore distinguish that the performance improvement brought by RHDP is not limited to the reward-function design alone, but stems from the hacking-detection and penalization mechanism. The LVoG constraint prevents reward over-optimization via accurate cheating detection. The forward KL constraint preserves language fluency from the SFT model. Together, these constraints effectively mitigate output distribution collapse and achieve superior overall performance.

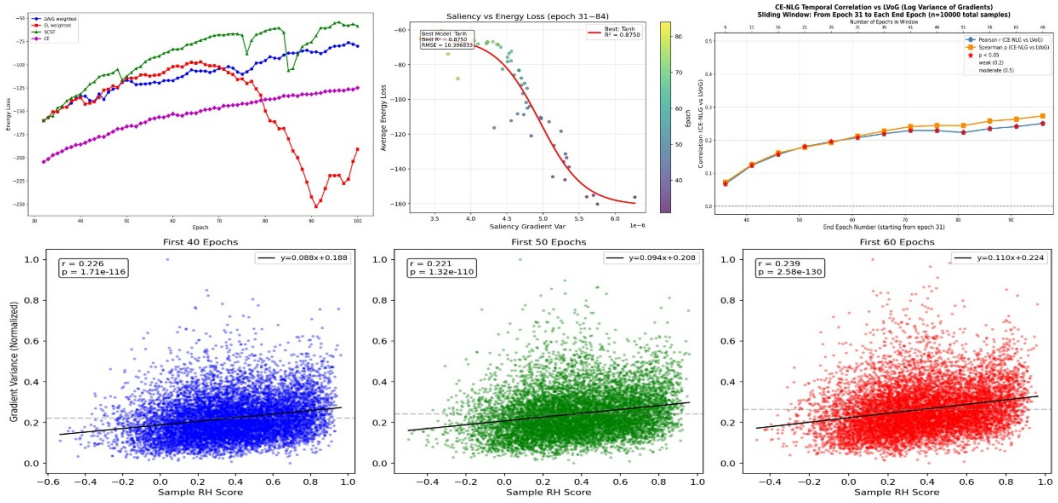

Figure 3: (a) The EL changes on the validation set of the model trained with CE, the model optimized with SCST, the model constrained with gradient variance, and the model constrained with energy loss. (b) The relationship between EL and VoG at the validation set level. (c) The stability of the Pearson and Spearman correlation coefficients of the proposed LVoG for hacking indication. (d)–(f) The relationship between the sample-level reward correlation and LVoG at different training stages.

4.5 Reward Hacking Risk Assessment

Evaluation of the reward-hacking (RH) detector. During training, moderate energy-loss increases and gradient-variance decreases are normal—reflecting steadily improving generalization—but excessive changes indicate hacking. As shown in Fig. 3a, both LVoG and EL suppress excessive energy-loss increases. Following (Laidlaw et al., 2025), we quantify hacking as the correlation between proxy reward (weighted NLG) and gold reward (the original RadCliQ)—lower correlation means stronger hacking, so at the sample level LVoG is negatively correlated with hacking degree (Fig. 3(d)–(f)). When shortcuts occur, last-layer energy loss rises excessively while the gradient first drops sharply then stabilizes, with both correlations significant (Fig. 3c). EL and LVoG are coupled forward- and backward-view observations of the same representation degradation and hence exhibit an inherent near-quadratic relation (Fig. 3b); owing to upstream Jacobian statistics, advantage computation, and sample-scale differences, however, only a significant monotonic rather than an exact functional relation can be expected. Notably, EL shows clear hacking signals only in the last layer, whereas LVoG's indication strengthens over time because its gradient is computed over the entire model. EL also fluctuates more and yields lower gains than LVoG (Table 2).

RH risk assessment of different reward functions. We analyze the loopholes of mainstream reward functions—weighted NLG (Yi et al., 2025), RadGraph F1 (Miura et al., 2021), BERTScore (Zhang et al., 2020), the original RadCliQ (Ori) based on CheXbert cosine similarity (Yu et al., 2023), and the improved RadCliQ (Ours)—by optimizing the base model with each and comparing PR values over 14 diseases, fluency and clinical accuracy, PR balance, length anomaly scores, and output entropy (Fig. 4). We also project validation-set predictions into a latent space using MediBioDeBERTa (Kim et al., 2023) and visualize them via t-SNE (van der Maaten & Hinton, 2008) (Fig. 5); RL fine-tuning shifts the distribution, with outliers indicating hacking. All metrics are normalized (larger is better). Weighted NLG and BERTScore cannot directly optimize clinical accuracy and thus produce conservative templates with limited clinical gains. RadGraph F1 and CheXbert-based rewards improve clinical detection but suffer severe length anomalies, repetition, and collapsed output distributions (low entropy). The improved RadCliQ balances fluency and clinical accuracy; although its detection on minority diseases remains poor (the inherent imperfection of verifiable rewards), it keeps the RL distribution close to the SFT model. With LVoG and forward KL constraints, outliers are further reduced, whereas EL causes a large distribution shift. The common implicit reward-manipulation behaviors these functions induce—template collapse, length exploitation, and repetitive generation—are illustrated in Appendix G.

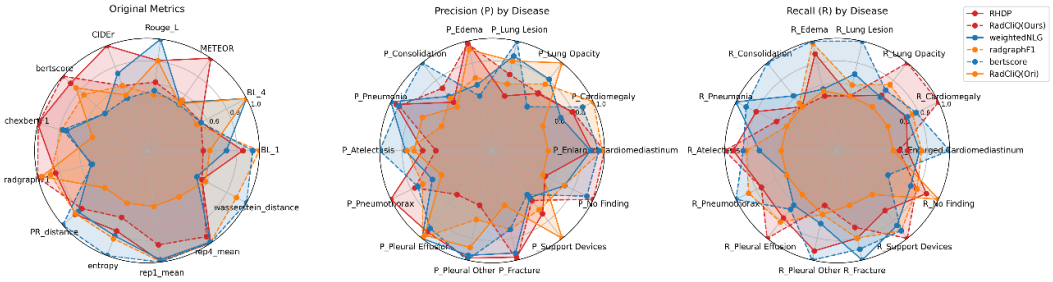

Figure 4: The radar charts of the model performance with different reward functions. (a) The common NLG metrics, CE metrics, and the RH evaluation metrics used in this work. (b) The precision of the 14 common diseases. (c) The recall of the 14 common diseases.

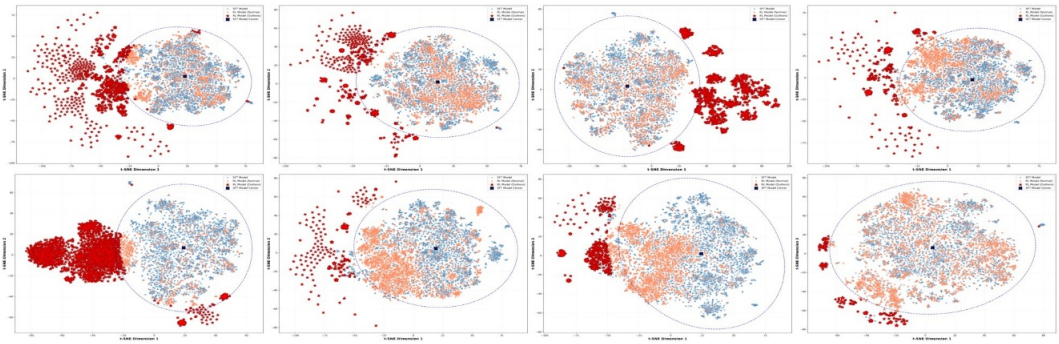

Figure 5: The t-SNE visualization of the validation sets of different models in the feature space. The first row from left to right shows the models based on weighted NLG, BERTScore, RadGraph F1, and RadCliQ (Ori) rewards, and the second row shows the models with the EL constraint, RadCliQ (Ours), the LVoG constraint, and the KL divergence based on hacking detector.

RH risk assessment of different divergence modulation strategies. A sharp output distribution is the most direct signature of reward hacking, so keeping RL fine-tuning close to the SFT model has become a consensus (Li et al., 2025). We compare reverse KL, forward KL, Jensen–Shannon (JS) divergence, and our unified hacking-aware modulation. Reverse KL forces mode collapse, JS gives moderate stability at a performance cost, and forward KL over-constrains all samples indiscriminately. In contrast, our modulation enables stable training, sustained reward growth, and restored proxy–gold correlation, effectively alleviating hacking (Fig. 6).

4.6 Qualitative Analysis

Fig. 7 presents example generations from ablation experiments and prior work. LHR-RFL and R2GenRL fail to identify the removed catheter and hallucinate substantially despite correctly diagnosing pleural effusion and lung consolidation. Our RHDP identifies both conditions while reducing irrelevant output, demonstrating improved clinical accuracy and hallucination mitigation. However, all models remain inaccurate on extra-thoracic devices and intra-thoracic catheters, requiring more precise external knowledge.

4.7 Generalization and Further Analyses

To further examine reliability and generality, we report additional analyses in the appendix. Appendix A connects EL and LVoG through forward–backward coupling and decomposes LVoG into CE–RL transition, reward-side advantage, and policy-side sensitivity dynamics. RHDP is then evaluated across RRG architectures (Appendix B), model scales and Qwen2 VLMs with PPO (Appendix C), attention distributions (Appendix D), hyperparameters (Appendix E), alternative detectors and occupancy measures (Appendix F), and deliberate hacking induction (Appendix G).

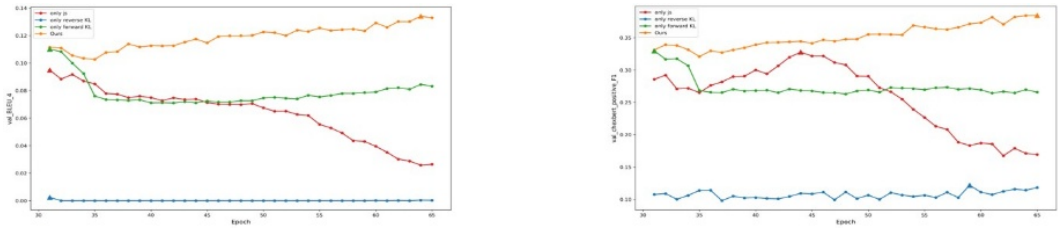

Figure 6: The training dynamics of the models with different divergence modulation strategies. (a) The BLEU-4 metric on the validation set. (b) The CheXbert F1 metric on the validation set.

Image	GT	Base	Base +LVoG	Base+Consis
	as compared to the previous exam the patient has been extubated and the nasogastric tube has been removed . the extent of the pre-existing pleural effusions have bilaterally increased . there is <unk> cardiomegaly with bilateral extensive areas of atelectasis . mild-to-moderate fluid overload . no focal parenchymal opacity suggest pneumonia .	right chest drain is in position since earlier study . there is a right chest tube with tip in the right atrium and right ventricle . small right pleural effusion is unchanged . there is moderate left-sided pleural effusion with blunting of the left costophrenic angle on the left . provided . there is no pneumothorax . comparison is made with interval placement of a right shoulder <unk>	heart size is normal . cardiomediastinal silhouette and hilar contours are been removed . right chest swan-ganz catheter are unchanged in appropriate position in the interval . small left pleural effusions are again demonstrated and bilateral pulmonary edema have increased retrocardiac opacities compatible with a small-to-moderate left pleural effusion . no pneumothorax is seen .	a right chest tube is in standard position . a left chest tube is in place . since the prior study there is a small right pleural effusion with adjacent atelectasis and small right pleural effusion . left lung is clear . heart is mildly enlarged . there is mild pulmonary edema . the right lung is clear .
	Base+KL	Ours RHDP	LHR-RFL	R2GenRL
	right picc line tip terminates in the mid svc . the cardiomediastinal silhouette is stable . there is mild perihilar prominence with vascular engorgement . there is worsening interstitial edema with severe pulmonary edema . there is bilateral small pleural effusions . there is no pneumothorax .	portable frontal chest radiograph demonstrates a left chest tube with course projecting over the right mid lung with a small right pleural effusion and left lower lung opacification consistent with atelectasis . no pneumothorax present . the cardiac silhouette is moderately enlarged for this in positioning . no focal consolidation pleural effusion or interstitial edema is present .	the patient has received a nasogastric tube with its tip <unk> to the gastroesophageal junction the gastric stomach . the tip of the tube is still in the stomach . the pleural course of the line is unchanged . there is unchanged evidence of well-expanded and <unk> complications at the valvular shows a small pleural effusions and no pneumothorax . minimal pulmonary edema is still present . the known osseous structures continues to the exception of the thoracic aorta the swan-ganz catheter is unchanged in appearance .	as compared to the previous radiograph the extent of a left pleural effusion has received a small catheter . the left pigtail catheter has been removed . the tip of the tube projects over the left internal gas probably in the lower distal course . no overt pulmonary edema . no new parenchymal opacities . normal size of the cardiac silhouette .

Figure 7: The quality comparison of the reports generated by different methods. The red text indicates the sentences that are clearly wrong compared with the ground truth, and the green text indicates the correct judgments.

5 Conclusion

We present RHDP, a reward-hacking detection and penalization framework for RL-fine-tuned radiology report generation. Its central statistic, LVoG, is the long-window variance of the joint-loss input gradient and serves as a composite sample-wise reliability indicator whose predictive power comes from fusing CE–RL transition, reward-side advantage, and policy-side sensitivity dynamics rather than advantage scaling alone. Across two benchmarks and multiple architectures, RHDP improves the clinical-accuracy–fluency trade-off and mitigates reward exploitation; however, LVoG calibration, the proxy–gold correlation analysis target, and our attention-collapse interpretation remain empirical and preliminary.

Acknowledgment

We thank the anonymous reviewers for their valuable feedback and suggestions. To preserve double-blind anonymity, funding sources and computing resources are omitted here and will be disclosed in the camera-ready version.

AI use statement

This paper was prepared with AI-based tools (including ChatGPT and DeepSeek), limited to non-scientific aspects of writing and production, with partial assistance in experimental-result interpretation. Specifically, we used large language models (LLMs) for:

- Language polishing and proofreading to improve readability and grammatical correctness, as we are non-native English speakers.
- Formatting assistance, including LaTeX code refinement, figure placement, and document structure organization.

- Translation support for converting initial drafts and experimental notes into coherent academic English.
- Experimental result interpretation and presentation. The discovery and mathematical formulation of gradient variance as a novel indicator were conducted entirely by the authors; AI assistance was limited to articulating the connection between this statistical measure and reward hacking.
- All AI-assisted content was reviewed, verified, and revised by the authors. The authors take full responsibility for the integrity of all scientific content in this paper.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics. We use the publicly available MIMIC-CXR and IU-Xray datasets, which have been de-identified and approved for research. No new patient data or human subjects were involved. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

The complete source code, training scripts, hyperparameters, and evaluation pipelines are submitted as a compressed archive in the supplementary materials. Detailed experimental settings are in Section 4.2, and data preprocessing steps are provided in the supplementary materials. The submitted code and documentation suffice to reproduce the reported results.

REFERENCES

Chirag Agarwal, Daniel D'Souza, and Sara Hooker. Estimating example difficulty using variance of gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

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

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. KiUT: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Baoyu Jing, Zeya Wang, and Eric Xing. 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 (ACL)*, pp. 6570–6580, 2019.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.

Eunhui Kim, Yuna Jeong, and Myung-Seok Choi. Medibiodeberta: Biomedical language model with continuous learning and intermediate fine-tuning. *IEEE Access*, 11:141036–141044, 2023.

Rohit Krishnan and Jon Evans. RLNVR: Reinforcement learning from non-verified real-world rewards. *arXiv preprint arXiv:2508.12165*, 2025.

Cassidy Laidlaw, Shivam Singhal, and Anca Dragan. Correlated proxies: A new definition and improved mitigation for reward hacking. In *International Conference on Learning Representations (ICLR)*, 2025.

Wangyu Lang, Zhi Liu, and Yijia Zhang. DACG: Dual attention and context guidance model for radiology report generation. *Medical Image Analysis*, 99, 2025.

Long Li, Zhijian Zhou, Jiaran Hao, Jason Klein Liu, Yanting Miao, Wei Pang, Xiaoyu Tan, Wei Chu, Zhe Wang, Shirui Pan, Chao Qu, and Yuan Qi. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward. *arXiv preprint arXiv:2509.07430*, 2025.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Yuchun Miao, Sen Zhang, Liang Ding, Yuqi Zhang, Lefei Zhang, and Dacheng Tao. The energy loss phenomenon in RLHF: A new perspective on mitigating reward hacking. In *International Conference on Machine Learning (ICML)*, 2025.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 448–458, 2022.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1179–1195, 2017.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations (ICLR)*, 2014.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Quan Tang, Liming Xu, Yongheng Wang, Bochuan Zheng, Jiancheng Lv, Xianhua Zeng, and Weisheng Li. Dual-modality visual feature flow for medical report generation. *Medical Image Analysis*, 101, 2025.

Tim Tanida, Philipp Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7433–7442, 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015.

Sharon Vera, Bar Yiftach, Kedar Eli, et al. Deep-learning natural language processing for oncological applications. *The Lancet Oncology*, 21(12):1553–1556, 2020.

Chaojie Wang, Haonan Shi, Long Tian, Bo An, and Shuicheng Yan. Removing prompt-template bias in reinforcement learning from human feedback. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Jun Wang, Abhir Bhalerao, Tianyang Yin, Simon See, and Yulan He. CAMANet: Class activation map guided attention network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Kequan Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Songtao Wang, Quang Hieu Pham, Fangcong Yin, Xinpeng Wang, Jocelyn Qiaochu Chen, Greg Durrett, and Xi Ye. Detecting and suppressing reward hacking with gradient fingerprints. *arXiv preprint arXiv:2604.16242*, 2026.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. METransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11558–11567, 2023.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Bo Yan, Mingtao Pei, Meng Zhao, Caifeng Shan, and Zhijun Tian. Prior guided transformer for accurate radiology reports generation. *IEEE Journal of Biomedical and Health Informatics*, 26 (11):5631–5640, 2022.

Xiulong Yi, You Fu, Rong Hua, Ruiqing Liu, and Hao Zhang. Unsupervised disease tags for automatic radiology report generation. *Biomedical Signal Processing and Control*, 89:105742, 2024.

Xiulong Yi, You Fu, Jianzhi Yu, Ruiqing Liu, Hao Zhang, and Rong Hua. LHR-RFL: Linear hybrid-reward-based reinforced focal learning for automatic radiology report generation. *IEEE Transactions on Medical Imaging*, 44(3):1494–1504, 2025.

Xiulong Yi, You Fu, Enxu Bi, Jianguo Liang, Hao Zhang, Jianzhi Yu, Qianqian Li, Rong Hua, and Rui Wang. Radiology report generation via visual-semantic ambivalence-aware network and focal self-critical sequence training. *Neural Networks*, 194, 2026.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andrew Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4 (9), 2023.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Hui Zhang, Peng Zeng, Lianli Gao, et al. SPT: Spatial pyramid transformer for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023a.

Jing Zhang, Yuan Xie, Weifeng Ding, et al. Cross on cross attention: Deep fusion transformer for image captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8): 4257–4268, 2023b.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020.

Tianyu Zhang, Tao Tang, Xin Wang, et al. RadioLOGIC, a healthcare model for processing electronic health records and decision-making in breast disease. *Cell Reports Medicine*, 4(8), 2023c.

Xuemiao Zhang and Junfei Liu. Adversarial training with comprehensive objective for medical image report generation. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2023.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. No free lunch: Rethinking internal feedback for LLM reasoning. *arXiv preprint arXiv:2506.17219*, 2025.

Appendix Overview.

- **A. Heuristic Analysis of LVoG and Reward-Hacking Risk** — forward–backward coupling between LVoG and the energy loss, the reward–policy decomposition of LVoG, the joint-degradation interpretation of reward hacking, and an optimization-level argument for LVoG-based reweighting.
- **B. Generalization Analysis** — applying RHDP to different RRG architectures (R2Gen, UDT, and DACG).
- **C. Model-Size Sensitivity Analysis** — transferring RHDP across model scales and to Qwen2 vision-language models.
- **D. Cross-Attention Distribution Evaluation** — attention collapse under reward exploitation and its mitigation.
- **E. Hyperparameter Tuning** — hyperparameters of the reward and divergence modulation strategies.
- **F. Limitations of Existing Methods and Empirical Choices of LVoG** — the instability of the EL-based detector, the limited effect of occupancy-measure regularization, and the empirical design of LVoG (MoG, VoG, and sliding-window VoG).
- **G. Reward-Hacking Induction Experiments** — verifying the reliability of LVoG by deliberately rewarding hacking behavior.

A HEURISTIC ANALYSIS OF LVOG AND REWARD-HACKING RISK

This section first analyzes the relation between Energy Loss (EL) and the proposed LVoG from a forward–backward coupling perspective of internal representations. We then decompose LVoG into proxy-advantage-driven optimization pressure and image-conditioned policy sensitivity, and separately examine their associations with sample-wise proxy–gold reward alignment. This decomposition is used to rule out the alternative explanation that the predictive signal of LVoG is predetermined solely by the algebraic scaling of the advantage. Finally, we provide an optimization-level argument for why LVoG-based reweighting favors more reliable policy updates and suppresses reward hacking. The central view is that **reward hacking is associated with a joint degradation of reward-side discriminability and policy-side dynamic plasticity, while the original LVoG uses advantage-weighted input-sensitivity dynamics to implicitly distinguish benign stable convergence from pathological stable collapse**.

A.1 FORWARD–BACKWARD COUPLING BETWEEN LVOG AND EL

For an input sample x, decompose the terminal representation pathway as

$$x \xrightarrow{F_\theta} z \xrightarrow{G_\theta} h, \tag{15}$$

where z and h denote the input and output of the final representation block. Let

$$J(x) = \frac{\partial h}{\partial z}, \qquad B(x) = \frac{\partial z}{\partial x} \tag{16}$$

be the local Jacobians. EL is a zeroth-order forward activation statistic,

$$\Delta E(x) = \|z\|_1 - \|h\|_1. \tag{17}$$

Under a local linearization $h \approx Jz$, a larger ΔE corresponds to stronger contraction of the representation norm. In the backward direction, letting $\delta = \nabla_h \mathcal{L}$ denote the terminal back-propagated signal, the input gradient satisfies

$$\nabla_x \mathcal{L} = B^\top J^\top \delta. \tag{18}$$

Thus, forward representation propagation is governed by J, whereas backward sensitivity propagation is governed by $J^\top$. These two operators have the same singular-value spectrum. If, over a local training interval, the second-order statistics of δ and B are approximately stationary and isotropic in the effective subspace,

$$\mathrm{Cov}(\delta) \approx \sigma_\delta^2 I, \qquad BB^\top \approx \beta I, \tag{19}$$

then

$$\mathbb{E}\|\nabla_x \mathcal{L}\|_2^2 = \sigma_\delta^2 \,\mathrm{tr}(JBB^\top J^\top) \approx \sigma_\delta^2 \beta \|J\|_F^2. \tag{20}$$

If additionally $\mathrm{Cov}(z) \approx \sigma_z^2 I$, then

$$\mathbb{E}\|h\|_2^2 = \mathbb{E}\|Jz\|_2^2 \approx \sigma_z^2 \|J\|_F^2. \tag{21}$$

Hence, forward representation energy and backward input sensitivity share the same local spectral quantity $\|J\|_F^2$. EL measures forward representation retention, whereas LVoG measures the temporal dynamics of the corresponding backward sensitivity channel. They are therefore not identical statistics, but can be interpreted as coupled forward and backward observations of the same degradation of input-conditioned representation retention under fixed scaling and local stationarity assumptions.

For further intuition, approximate the local mapping as $J \approx cR$, where R is approximately norm preserving and $c > 0$ denotes a sample-wise representation-retention factor. Then

$$\|h\|_1 \approx c\|z\|_1, \qquad \Delta E \approx \|z\|_1(1-c), \tag{22}$$

which gives

$$c \approx 1 - \frac{\Delta E}{\|z\|_1}. \tag{23}$$

Since a second-order backward-sensitivity statistic is locally of order c^2, we obtain the heuristic relation

$$\mathrm{LVoG} \propto c^2 \approx \left(1 - \frac{\Delta E}{\|z\|_1}\right)^2. \tag{24}$$

This predicts a generally negative and nonlinear EL–LVoG association, which is consistent with the empirical observation in Fig. 8. The derivation is intended to expose a shared local representation-retention factor rather than to claim global equivalence.

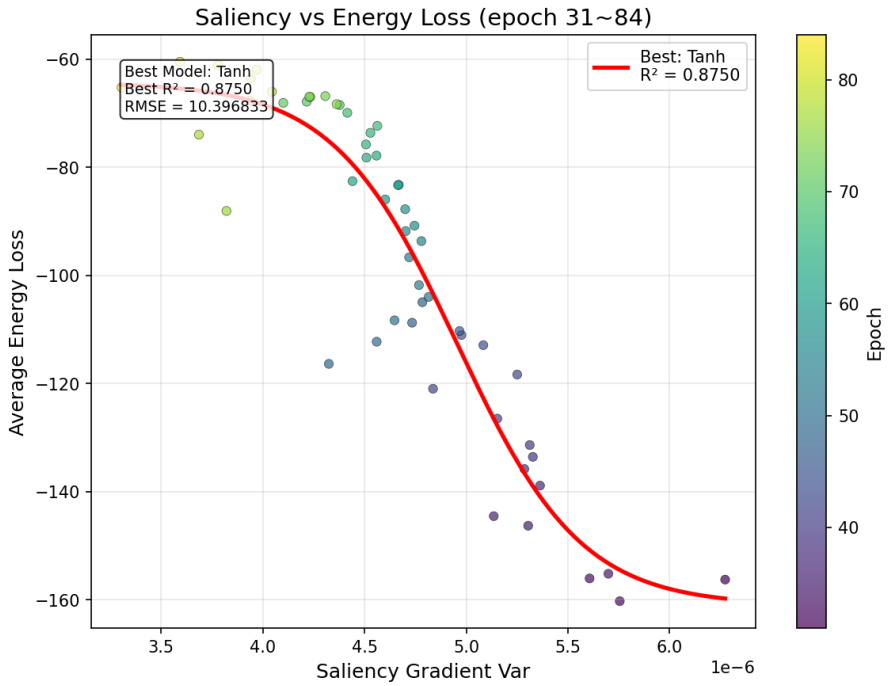

Figure 8: EL and LVoG show a strong nonlinear association, offering an intuitive explanation of their complementary roles—capturing representational collapse from the forward and backward perspectives, respectively.

A.2 Decomposing Joint-Loss LVoG Into CE-, Reward-, and Policy-Side Dynamics

For sample i at training step t, let $y^s_{i,t}$ be the sampled sequence, $y^b_{i,t}$ the baseline sequence, and r^p the proxy reward. The SCST proxy advantage is

$$A^p_{i,t} = r^p(y^s_{i,t}) - r^p(y^b_{i,t}). \tag{25}$$

Define the advantage-free policy negative log-likelihood

$$\ell^{\mathrm{pol}}_{i,t} = -\frac{1}{T_{i,t}} \sum_{k=1}^{T_{i,t}} \log \pi_{\theta_t}(y^s_{i,t,k} \mid y^s_{i,t,<k}, x_i). \tag{26}$$

With stop-gradient through the advantage, the RL term and the joint CE–RL objective used by RHDP are

$$\mathcal{L}^{\mathrm{RL}}_{i,t} = \mathrm{sg}\big(A^p_{i,t}\big)\, \ell^{\mathrm{pol}}_{i,t}, \qquad \mathcal{L}^{J}_{i,t} = \alpha_1 \mathcal{L}^{\mathrm{CE}}_{i,t} + \alpha_2 \mathcal{L}^{\mathrm{RL}}_{i,t}. \tag{27}$$

Let $g^{\mathrm{CE}}_{i,t} = \nabla_{x_i} \mathcal{L}^{\mathrm{CE}}_{i,t}$ and $g^{\mathrm{pol}}_{i,t} = \nabla_{x_i} \ell^{\mathrm{pol}}_{i,t}$. The joint input gradient has the exact vector decomposition

$$g^{J}_{i,t} := \nabla_{x_i} \mathcal{L}^{J}_{i,t} = \alpha_1 g^{\mathrm{CE}}_{i,t} + \alpha_2 A^p_{i,t} g^{\mathrm{pol}}_{i,t}. \tag{28}$$

For the scalarization $\phi(g) = M^{-1}\|g\|_1$, the operational statistic is

$$G^{J}_{i,t} = \phi\big(g^{J}_{i,t}\big), \qquad \mathrm{LVoG}_{i,t} = \mathrm{Var}_{\tau \in \mathcal{W}_t}\left(G^{J}_{i,\tau}\right). \tag{29}$$

The above equation is exact at the vector level, but the scalarized joint gradient is not additively separable because ϕ contains an L_1 norm. The reverse and ordinary triangle inequalities give

$$\left|\alpha_1 \phi\big(g^{\mathrm{CE}}_{i,t}\big) - \alpha_2 \left|A^p_{i,t}\right| \phi\Big(g^{\mathrm{pol}}_{i,t}\Big)\right| \le G^{J}_{i,t} \le \alpha_1 \phi\big(g^{\mathrm{CE}}_{i,t}\big) + \alpha_2 \left|A^p_{i,t}\right| \phi\Big(g^{\mathrm{pol}}_{i,t}\Big). \tag{30}$$

Therefore, the original LVoG is neither the product nor the sum of independent component variances. It can vary with CE-side magnitude, reward-scaled policy magnitude, and the changing directional agreement or cancellation between the CE and RL gradients. This cross-objective interaction is part of the statistic rather than an error term that can be discarded.

We use four auxiliary diagnostics to identify the roles of these channels. First, for steps satisfying $\left|A^p_{i,t}\right| > \varepsilon$, the advantage-normalized joint-gradient variance is

$$\mathrm{nLVoG}_{i,t} = \mathrm{Var}_{\tau\in\mathcal{W}_t}\left[\phi\left(\frac{g^J_{i,\tau}}{A^p_{i,\tau}}\right)\right], \qquad \frac{g^J_{i,t}}{A^p_{i,t}} = \alpha_1\frac{g^{\mathrm{CE}}_{i,t}}{A^p_{i,t}} + \alpha_2 g^{\mathrm{pol}}_{i,t}. \tag{31}$$

Near-zero advantages must be excluded or stabilized with a reported ε. Importantly, nLVoG is not a pure policy statistic: empirically it is dominated by the g^{CE}/A^p term and therefore mainly reflects the competition between the CE anchor and the reward-scaled RL update. The CE-only long-window variance is

$$\mathrm{cLVoG}_{i,t} = \mathrm{Var}_{\tau\in\mathcal{W}_t}\left[\phi\big(g^{\mathrm{CE}}_{i,\tau}\big)\right]. \tag{32}$$

The pure RL policy-sensitivity variance, which removes instantaneous advantage scaling and excludes the CE term, is

$$\mathrm{pLVoG}_{i,t} = \mathrm{Var}_{\tau\in\mathcal{W}_t}\left[\phi\Big(g^{\mathrm{pol}}_{i,\tau}\Big)\right]. \tag{33}$$

Finally, the reward-side variance of the gradient-scaling magnitude is

$$\mathrm{VoA}_{i,t} = \mathrm{Var}_{\tau\in\mathcal{W}_t}\left(\left|A^p_{i,\tau}\right|\right). \tag{34}$$

We quantify sample-wise optimization consistency by the proxy–gold reward alignment

$$C_{i,t} = \mathrm{Corr}_{\tau\in\mathcal{W}_t}\left(\Delta r^p_{i,\tau}, \Delta r^g_{i,\tau}\right), \qquad H_{i,t} = 1 - C_{i,t}, \tag{35}$$

where larger C indicates stronger alignment and larger H indicates more severe misalignment. Consequently, a positive LVoG–C association is exactly the same directional finding as a negative LVoG–H association. The scatter plots appear positive because their horizontal axis is the alignment score C, not the misalignment score H. This sign convention is used consistently throughout the paper.

The component analyses at fixed training windows (Rows 1–4 of Fig. 9) reveal distinct, complementary roles. The original LVoG has the strongest positive association with the sample-level proxy–gold alignment C (equivalently, the strongest negative association with the hacking risk H). VoA, the reward-side variance of the advantage magnitude, follows the same monotonic direction and provides reward discriminability information: severe misalignment is accompanied by a more homogeneous, saturated, or weakly discriminative advantage process. By contrast, both the advantage-normalized joint-gradient variance (nLVoG, dominated by the CE/$|A|$ term) and the CE-gradient variance (cLVoG) exhibit a Gaussian-shaped inverted-U relationship with C: they become large in a restructuring or transition region but are small at both stable endpoints, so they localize transitions without providing reliable directionality. pLVoG, the variance of the pure image-conditioned RL policy sensitivity, remains positively associated with C (equivalently, negatively associated with H), although more weakly than the original LVoG. This establishes that contraction of pure policy sensitivity is itself a reward-hacking correlate, yet no isolated component reproduces the discriminative strength of the joint statistic.

The fifth row of Fig. 9 adds a temporal dimension by tracing how the association strength of each statistic with C evolves as the training window lengthens. LVoG and VoA behave consistently: their positive association with C strengthens monotonically with training, indicating that the composite statistic and the reward-side channel become progressively more discriminative as misalignment develops. pLVoG behaves differently—its association with C first increases during the early-to-middle training stages and then declines in later epochs, tracing a pronounced arch-shaped trajectory. This dissociation is mechanistically informative: pLVoG measures image-conditioned policy plasticity and can rank samples only while the policy is still being restructured; once training drives samples into the pathological attractor, both benignly stable and shortcut-collapsed samples exhibit small pLVoG, and its ability to discriminate by alignment erodes. The joint statistic is immune to this erosion because the advantage-modulated gradient retains reward-side and cross-gradient interaction information even after policy-side sensitivity alone has saturated. Row 5 therefore shows that the superiority of LVoG is not merely cross-sectional but also temporal: its indication effectiveness persists and even improves, whereas that of any single channel is time-localized.

These findings determine the precise innovation of LVoG. The novelty is not the generic use of a gradient magnitude and not an exact algebraic factorization into pLVoG and VoA. Rather, LVoG uses the long-window variance of an end-to-end joint-objective input gradient as an online sample-wise reliability statistic. It fuses a CE–RL transition signal with two directional indicators—reward-side advantage discriminability and RL policy-side image-conditioned sensitivity—and preserves their cross-gradient interaction. Empirically, this fusion yields the strongest association with proxy–gold alignment among all components at every examined window and, uniquely, an association that keeps strengthening over training, making the original LVoG the appropriate statistic for training-time detection and modulation.

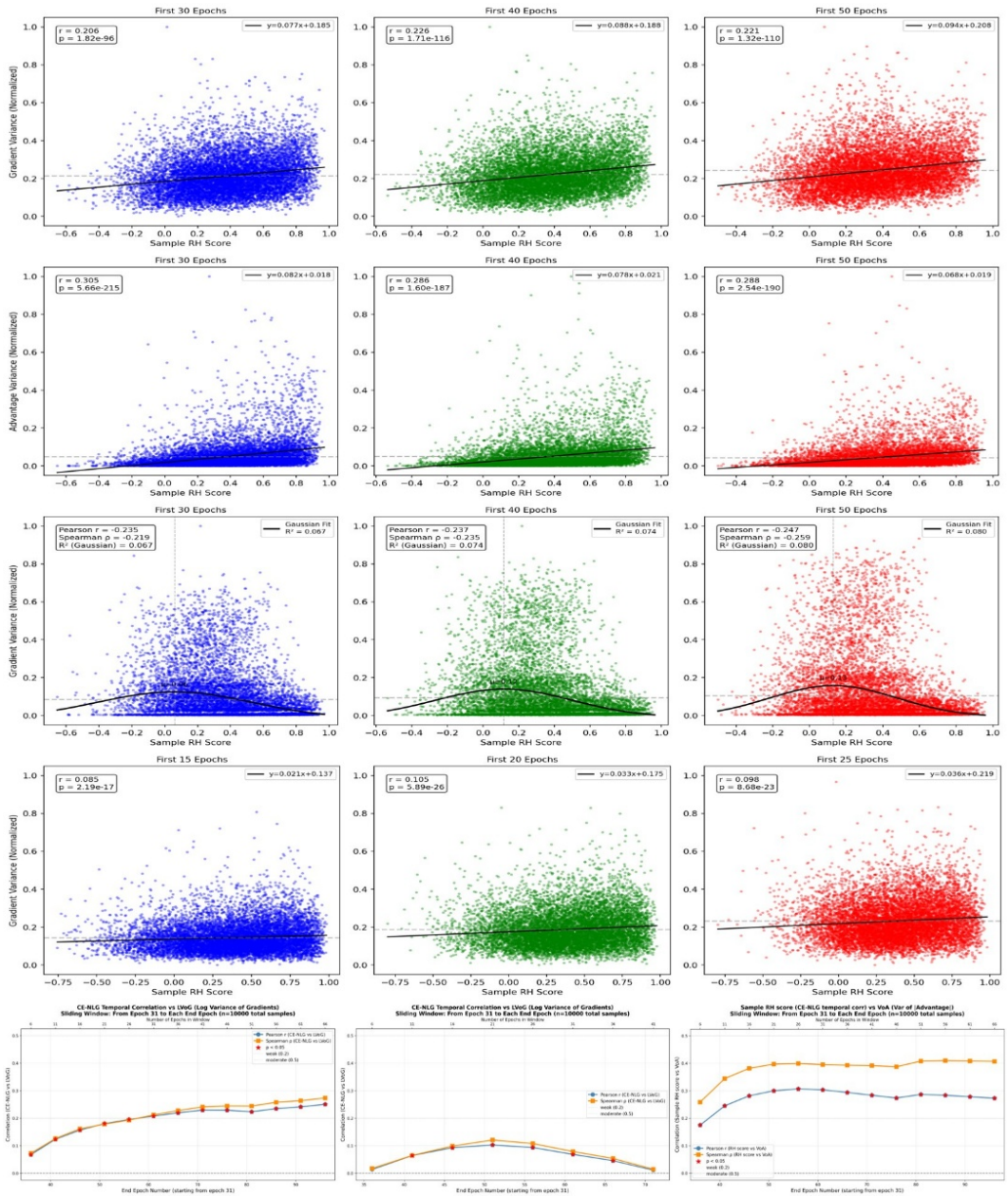

Figure 9: Five-row decomposition of the relationships between LVoG-related statistics and sample-level proxy–gold alignment at 30, 40, and 50 epochs. Row 1: original joint-loss LVoG, which is significantly positively associated with alignment (equivalently, negatively associated with hacking risk). Row 2: advantage variance (VoA), which shows the same monotonic direction and provides reward-side discrimination. Row 3: advantage-normalized joint-gradient variance (nLVoG), which is dominated by the CE/$|A|$ term and exhibits an inverted-U-shaped relationship; the CE-gradient variance (cLVoG) shows the same qualitative pattern. Row 4: pure RL policy-sensitivity variance (pLVoG), which remains positively associated with alignment but more weakly than the original LVoG. Together, the first four rows support the interpretation of LVoG as a composite rather than a single-channel indicator. Row 5: The variation of reward-optimization mismatch indication effectiveness of LVoG, pLVoG, and VoA across epochs.

A.3 Joint Degradation and the Risk-Indication Mechanism of LVoG

The above decomposition suggests three heuristic regimes during SCST optimization. First, under **benign stable convergence**, the policy has formed a stable generation mode that remains consistent with the image; pLVoG can therefore be low, while proxy–gold alignment remains high and reward-side discrimination does not exhibit pathological collapse. Second, under a **transitional or restructuring regime**, the model actively changes its image dependence and generation strategy, giving high pLVoG and strong dynamic plasticity. Third, under **pathological stable collapse**, the model has entered a reward-favored shortcut attractor; image-conditioned policy dynamics no longer change meaningfully, and the advantage signal becomes homogeneous or weakly varying, yielding low pLVoG, low VoA, and low original LVoG. Reward hacking is therefore not a single-variable "low-gradient-variance" event, but a joint degradation of the reward-discrimination channel and the policy-adaptation channel.

This interpretation also explains why the original LVoG is more suitable than pLVoG alone as a practical detector and modulation variable. pLVoG localizes the region of strongest policy transition but cannot by itself distinguish the two low-variance endpoints. VoA contributes reward-side directionality. The joint-gradient statistic LVoG retains both sources of information and their cross-objective interaction in a single quantity, allowing its association with proxy–gold alignment to be more stable and closer to monotonic than that of pLVoG alone. Accordingly, pLVoG is primarily a mechanism-analysis statistic, VoA is a reward-discriminability dynamics statistic, and the original LVoG is the composite reliability indicator used for training-time control. This observation exactly motivates our joint modulation of the forward KL divergence using both LVoG and the immediate reward.

A.4 Why LVoG-Based Reweighting Promotes More Reliable Policy Optimization

Let the proxy policy gradient of sample i be

$$g_i^p = \mathbb{E}_{y\sim\pi_\theta}\left[A_i^p \nabla_\theta \log \pi_\theta(y \mid x_i)\right], \tag{36}$$

and let g_i^g denote the ideal gradient associated with the gold objective. Let $q_i \in [0,1]$ be the normalized LVoG reliability score and let $w(q_i)$ be monotonically increasing in q_i. The induced effective training distribution is

$$\tilde{p}(i) = \frac{p(i)w(q_i)}{\mathbb{E}_p[w(q)]}. \tag{37}$$

Let the sample-wise misalignment risk u_i be a monotone decreasing function of q_i. Since w increases with reliability and therefore decreases with risk,

$$\mathrm{Cov}_p\big(u, w(q)\big) \le 0. \tag{38}$$

Consequently,

$$\mathbb{E}_{\tilde{p}}[u] = \frac{\mathbb{E}_p[uw]}{\mathbb{E}_p[w]} = \mathbb{E}_p[u] + \frac{\mathrm{Cov}_p(u,w)}{\mathbb{E}_p[w]} \le \mathbb{E}_p[u]. \tag{39}$$

Thus, LVoG-based reweighting lowers the average misalignment risk of the effective sample distribution that actually contributes to policy updates.

Assume further that the sample-wise discrepancy between proxy and gold gradients is controlled by the risk score, i.e., for some $K > 0$,

$$\|g_i^p - g_i^g\| \le K u_i. \tag{40}$$

Then the bias of the weighted average update satisfies

$$\|\mathbb{E}_{\tilde{p}}[g^p - g^g]\| \le K\mathbb{E}_{\tilde{p}}[u] \le K\mathbb{E}_p[u]. \tag{41}$$

Therefore, provided that LVoG ranks sample reliability with nontrivial statistical accuracy, its use as a weight reduces an upper bound on the deviation of proxy-driven updates from the gold direction. This argument does not require LVoG to be an unbiased estimator of the true reward; it only requires LVoG to reliably rank more and less trustworthy sample-wise updates. Our experimental results have already demonstrated that LVoG serves as such a reliable sample ranker.

Summary. The analysis provides a unified interpretation. EL and LVoG observe input-conditioned degradation from the forward-representation and backward-gradient perspectives, respectively. pLVoG exposes the dynamic plasticity of the image-conditioned policy and reveals a non-monotonic structure consisting of benign stability, transitional adaptation, and pathological stability. VoA captures reward-side discriminability dynamics and disambiguates the two low-pLVoG endpoints. The original LVoG, through the joint-gradient variance $\mathrm{Var}(G^J)$, combines both channels and their cross-objective interaction into a reward-conditioned gradient-dynamics reliability indicator. We therefore interpret reward hacking as a joint degradation of reward-side and policy-side dynamics, rather than as low gradient variance alone, and use LVoG to rank and reweight policy updates toward regions with lower proxy–gold misalignment risk.

B Generalization Analysis

Our RHDP framework improves the performance of existing RRG models, including R2Gen (Chen et al., 2020), UDT (Yi et al., 2024), and DACG (Lang et al., 2025), as shown in Table 3. RHDP yields clear average gains on most configurations; on two of them (UDT on IU-Xray and R2Gen on MIMIC-CXR-10k), the average lexical shift is slightly negative because penalizing shortcut solutions suppresses repetitive n-gram template generation and therefore lowers rigid surface-overlap scores such as BLEU and ROUGE-L. This lexical drop is the expected cost of the same anti-hacking mechanism that prevents reward over-optimization in the main experiments, rather than a failure of generalization. Overall, this demonstrates the generalizability of our approach: its core lies not in modifying the model architecture but in leveraging LVoG to focus on reward-hacking (RH) samples.

Table 3: Performance comparison of several existing models with and without RHDP. "Avg.Δ" denotes the average relative improvement (%) of each configuration over the preceding one, and the best result within each base model is in bold.

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	R-L	CIDEr	Avg.Δ (%)
IU-Xray	R2Gen	0.458	0.300	0.219	0.169	0.186	0.366	**0.649**	–
	R2Gen+SCST	**0.533**	0.367	0.266	0.196	0.218	0.370	0.396	7.92
	R2Gen+SCST+RHDP	0.530	**0.373**	**0.268**	**0.197**	**0.225**	**0.372**	0.428	2.02
	UDT	0.441	0.279	0.196	0.147	0.180	0.349	**0.449**	–
	UDT+SCST	**0.500**	0.343	0.251	**0.190**	**0.215**	**0.410**	0.342	15.24
	UDT+SCST+RHDP	0.499	**0.352**	**0.254**	0.186	0.207	0.377	0.375	-0.08
	DACG	0.443	0.274	0.186	0.133	0.176	0.341	**0.366**	–
	DACG+SCST	0.494	0.331	0.238	0.177	0.204	0.366	0.349	15.99
	DACG+SCST+RHDP	**0.522**	**0.358**	**0.259**	**0.189**	**0.213**	**0.367**	0.348	4.83
MIMIC-CXR-10k	R2Gen	0.294	0.183	0.121	0.087	0.114	0.264	0.109	–
	R2Gen+SCST	**0.331**	0.207	0.143	**0.105**	0.123	**0.292**	**0.116**	12.78
	R2Gen+SCST+RHDP	**0.331**	**0.209**	**0.145**	0.104	**0.124**	0.283	0.109	-1.09
	UDT	0.304	0.182	0.117	0.082	0.101	0.250	0.077	–
	UDT+SCST	0.318	0.198	0.132	0.094	0.121	**0.276**	0.090	12.56
	UDT+SCST+RHDP	**0.330**	**0.205**	**0.139**	**0.101**	**0.124**	**0.276**	**0.106**	6.23
	DACG	0.261	0.164	0.108	0.074	0.107	0.257	0.074	–
	DACG+SCST	**0.300**	**0.187**	**0.128**	0.092	**0.114**	0.283	0.085	14.76
	DACG+SCST+RHDP	0.281	0.181	0.127	**0.094**	**0.114**	**0.287**	**0.100**	1.56

C Model-Size Sensitivity Analysis

Our proposed RHDP method consistently mitigates reward hacking across SCST-fine-tuned RRG models with fewer than 1B parameters (e.g., ViT, R2Gen, UDT (Yi et al., 2024), and DACG (Lang et al., 2025)), as shown in Appendix B. To assess its broader applicability, we further evaluate Qwen2 vision-language models (VLMs) trained with PPO (Schulman et al., 2017; Wang et al., 2024b), which also yield stable gains and effective hacking alleviation (Table 4). These results demonstrate that the proposed mechanism is not sensitive to the model scale or the specific RL algorithm.

Table 4: Validation on larger-scale VLM models (Qwen2-VL-2B).

Dataset	Method	BLEU-4	METEOR	ROUGE-L	CIDEr	BERTScore	CheXbert F1	Entropy
MIMIC-CXR-3k	CE	0.055	0.287	0.204	0.133	0.569	0.251	0.925
	RL_NLG	0.051	0.286	0.204	0.179	0.591	0.269	0.517
	RL_RadCliQ	0.034	0.259	0.197	0.116	0.555	0.279	0.680
	RL_RadCliQ+RHDP	0.052	0.264	0.199	0.210	0.568	0.294	0.630

D Cross-Attention Distribution Evaluation

We visualize the cross-attention maps of different methods in Fig. 10. Since RRG models exhibit multi-granularity image-text alignment and adjacent words share similar attention patterns, we aggregate maps over salient tokens for clarity (Wang et al., 2024a). R2GenRL (Chen et al., 2020) and LHR-RFL (Yi et al., 2025), both optimized with fluency-based rewards, suffer from severe attention collapse. This is manifested as sharpened distributions and misaligned visual grounding, driven by reward exploitation regardless of architectural complexity. In contrast, the improved RadCliQ reward with our hacking penalization effectively mitigates this collapse during RL training.

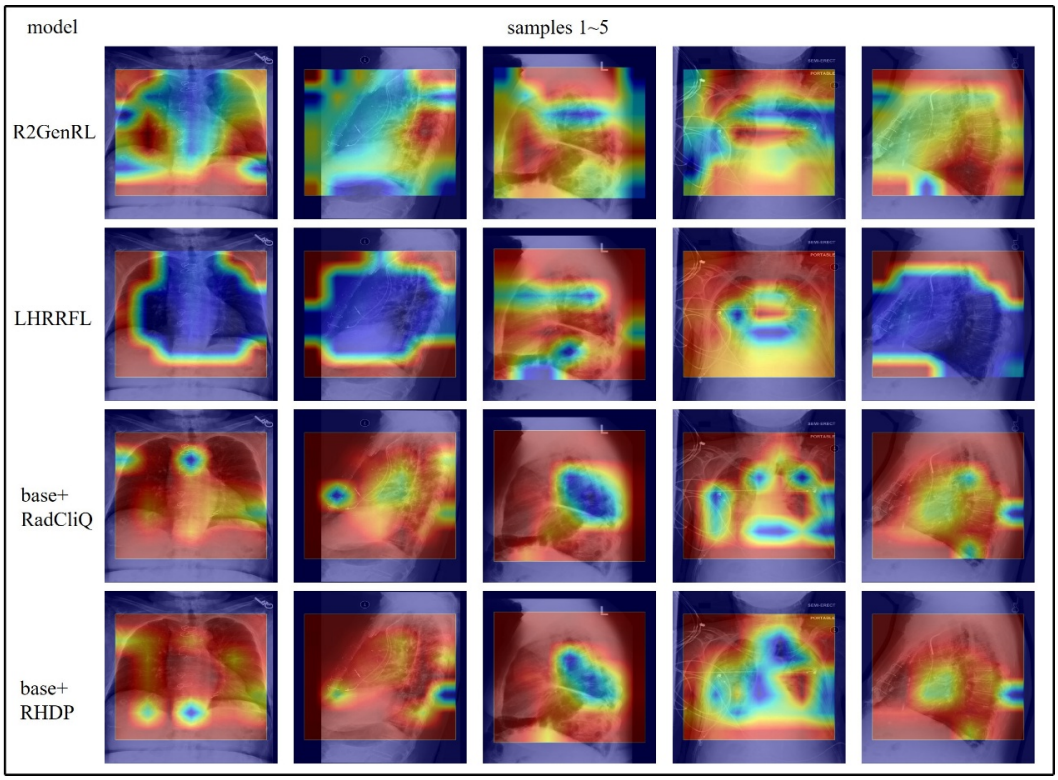

Figure 10: The cross-attention score maps of the top $k\%$ important words of different RL methods.

E Hyperparameter Tuning

We tune the hyperparameters of the two core components, as shown in Fig. 11, and obtain a set of optimal weights. An excessively high penalty on hacking behavior damages the ability of the model to learn templates. A reasonable divergence modulation can both preserve the high-performance sampling mode of the SFT model and retain the exploration capability of the model.

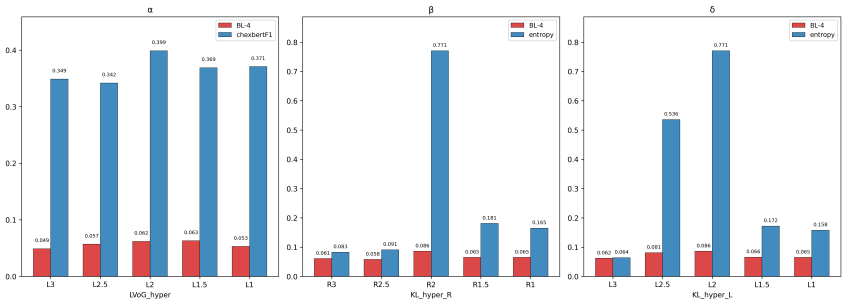

Figure 11: The hyperparameter tuning of the core components. (a) The penalization factor α of the reward modulation based on hacking behavior. (b) and (c) The penalization factors β and δ of the divergence modulation strategy based on the immediate reward and hacking behavior.

F Limitations of Existing Methods and Empirical Choices of LVoG

F.1 Limitations of the EL-based Detector

We compare the stability of EL (Miao et al., 2025) and LVoG as hacking indicators in Fig. 12. Although EL indicates hacking significantly for RRG models under general SCST training within a certain range of epochs, its indication is unstable: its significance fluctuates irregularly. Our LVoG, by contrast, remains continuously significant, benefiting from both the information richness brought by a longer window and the denoising ability obtained by aggregating over the entire model rather than a single layer.

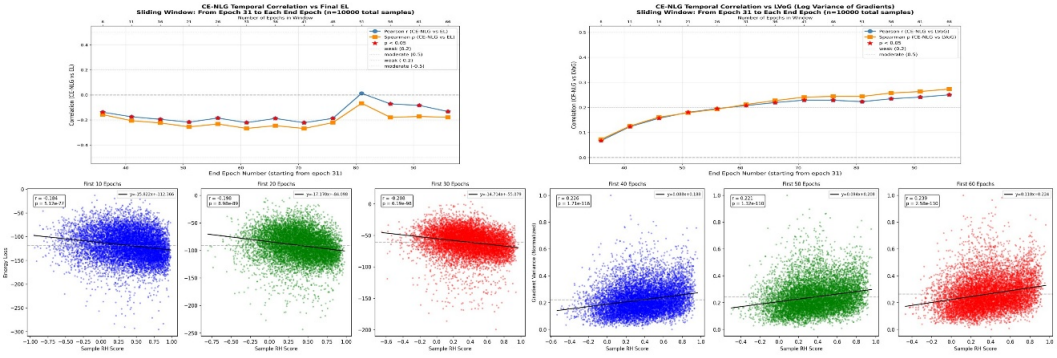

Figure 12: The top row shows the Pearson/Spearman correlations and their significance over the course of training for EL (left) and our LVoG (right); the latter is clearly more stable and thus superior. The bottom row shows the sample-level fitting results of the two statistics at individual epochs.

F.2 Limitations of Occupancy-Measure Regularization

Laidlaw et al. (2025) showed that RL fine-tuning breaks the correlation between the proxy reward and the gold reward on the state–action pairs known to the SFT model. They further proposed occupancy-measure regularization to mitigate this phenomenon; although it effectively restores the proxy–gold correlation in the common-probability region, its effect remains limited in low-probability regions. Following their methodology, we show in Fig. 13 how the reward correlation computed over all validation-set samples evolves across epochs. This correlation does not decrease with deeper RL fine-tuning; on the contrary, it becomes increasingly robust and stronger. This indicates that the drift induced by RL fine-tuning is a concealed, progressive reward hacking rather than a catastrophic, instantaneous collapse: the cumulative correlation of the whole training process decreases (Fig. 1), whereas the cross-sectional correlation within each epoch increases (Fig. 13).

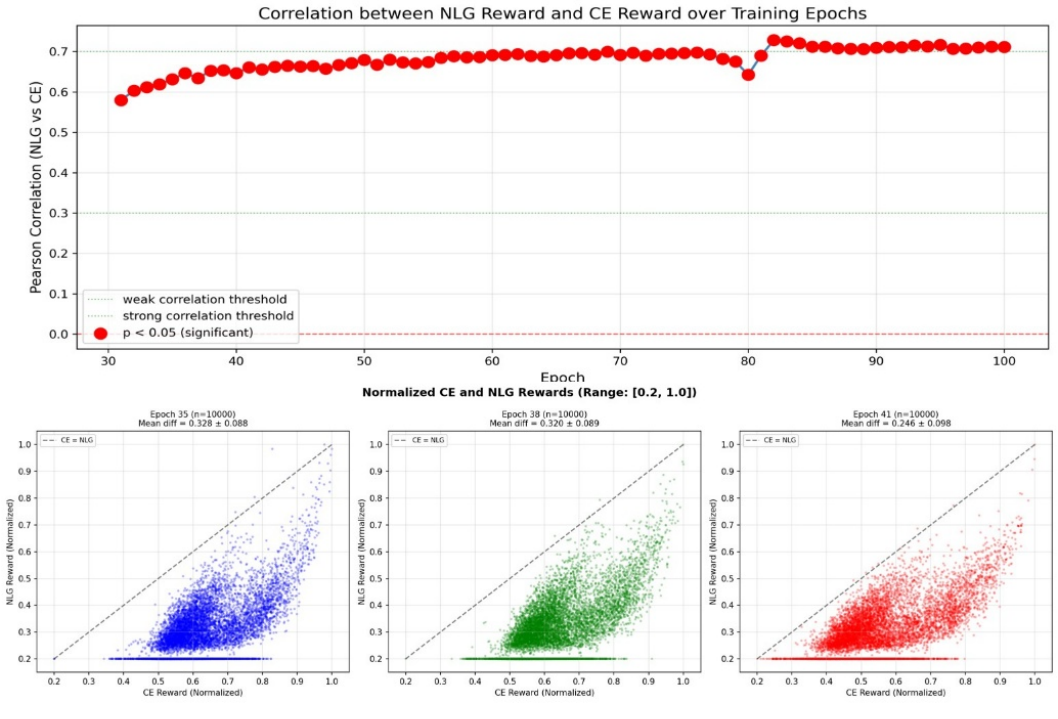

Figure 13: The within-epoch sample-level correlation between the proxy reward and the gold reward keeps increasing during RL fine-tuning. The top row shows the trend of this correlation over epochs, and the bottom row shows the sample-level fitting results at individual epochs.

We also use the cumulative output-probability distribution and the entropy as two alternative binning criteria for samples: low entropy indicates stable and determinate samples, while low cumulative probability indicates low-probability samples that the model seldom samples; high entropy indicates difficult samples about which the model is uncertain, while high cumulative probability indicates high-probability samples that the model samples easily. As shown in Fig. 14, the SFT model discriminates well across samples under both binning schemes, whereas RL fine-tuning destroys this discrimination. This further suggests that, in the SCST-fine-tuned RRG setting, judging sample-level hacking solely from the cross-sectional correlation at a single training step is limited; one should more broadly consider the historical dynamic information of each sample.

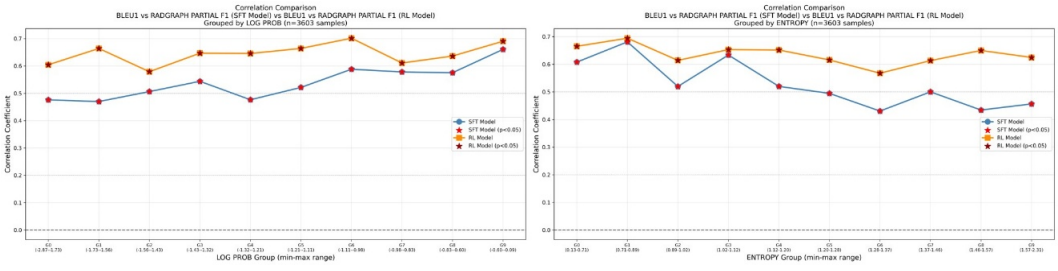

Figure 14: Using the cumulative output probability and the entropy as separate binning criteria, we compare how well these two variables discriminate the sample-level reward correlation of the SFT model and of the RL-fine-tuned model.

F.3 Empirical Design of the Proposed Statistic

The core innovation of this work is LVoG, a statistic based on the training dynamics of the model. Its design is intuition-driven: when a model begins to hack, it collapses rapidly and loses the ability

to understand samples, so the gradient of the loss with respect to the input x drops abnormally fast and then stabilizes. Because this statistic is empirical in nature, it is a limitation of this work; we therefore investigate several design choices below to better understand when and why it works.

F.3.1 MoG: The Mean of the Input Gradient over a Window of Length T

We first show that the mean of the input gradient over a window of length T, denoted MoG, also possesses a certain hacking-indication effect. Its computation, as well as its relation to VoG, is given as follows:

$$\mu_{n,t} = \frac{1}{M}\sum_{i=1}^{M} S_{n,t,i}, \quad n = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T \tag{42}$$

$$\mathrm{MoG}_n = \frac{1}{T}\sum_{t=1}^{T} \mu_{n,t}, \quad n = 1, 2, \ldots, N \tag{43}$$

$$\mathrm{VoG}_n = \mathrm{Var}(\mu_{n,t}), \quad n = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T \tag{44}$$

where M is the number of pixels, N is the number of samples, and T is the number of training epochs, which also serves as the window size. As shown in Fig. 15, MoG still indicates hacking with a certain significance, but its correlation with the hacking degree is much weaker than that of LVoG. This is because MoG, unlike LVoG, does not contain the second-order temporal information of a "sharp drop followed by stabilization"; nevertheless, it still correlates strongly with a short-window VoG (Fig. 16, left). This is determined by the property that the input gradient slowly decreases during training as the model gradually acquires generalization (Fig. 16, right).

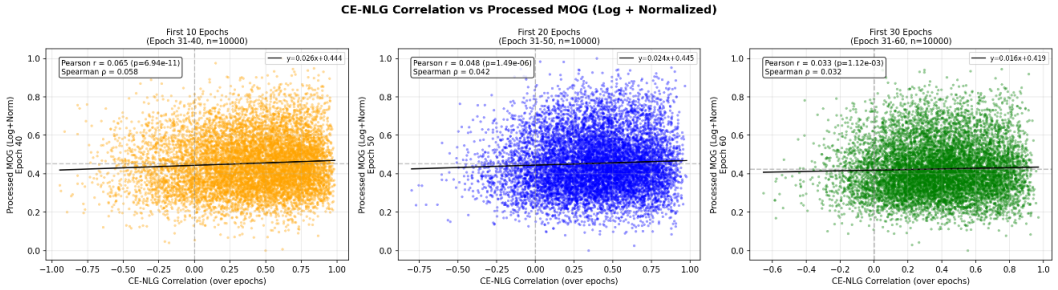

Figure 15: Significance analysis of the sample-level MoG–hacking-risk association at individual epochs.

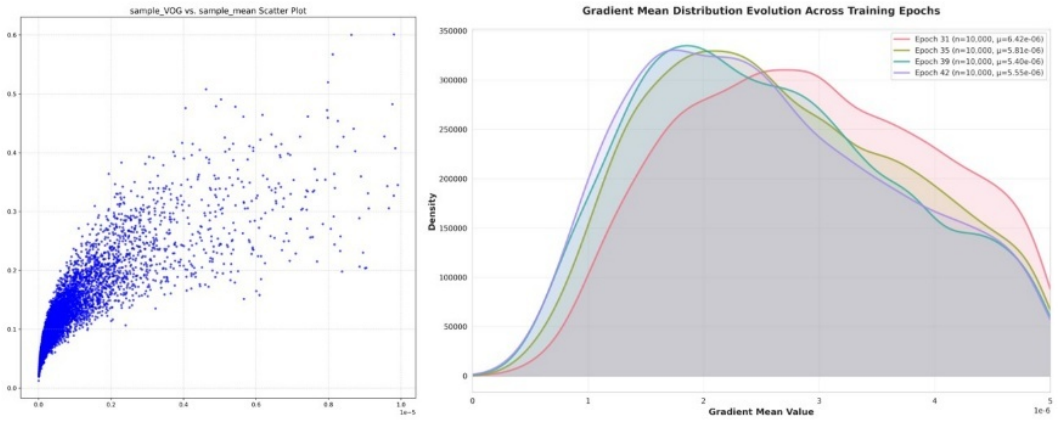

Figure 16: Left: the association between MoG and VoG. Right: the distribution of the mean input gradient, which shifts significantly to the left along with CE pre-training or RL fine-tuning.

F.3.2 SLIDING-WINDOW VOG

Note that "high LVoG corresponds to normal training whereas low LVoG corresponds to hacking" is not a universal law that is independent of the window position. If the window is short and happens to cover the rapid transition at the moment hacking occurs, the hacking trajectory may also temporarily exhibit a high variance. We therefore examine the sensitivity of the sliding-window VoG for hacking indication: we fix the penalty factor and choose different window widths K from 3 to 7, as shown in Table 5. Even a short-window VoG still improves model performance to some extent.

Table 5: Hyperparameter tuning for the window width K on the IU-Xray dataset; the reward function is weighted NLG.

K	α	BLEU-1	BLEU-4	METEOR	ROUGE-L	CIDEr
3	2	0.523	0.207	0.215	0.406	0.461
4	2	0.512	0.208	0.213	0.407	0.552
5	2	0.544	0.209	0.244	0.414	0.444
6	2	0.536	0.202	0.225	0.412	0.419
7	2	0.528	0.202	0.225	0.411	0.397
Base	–	0.483	0.177	0.197	0.378	0.393

G REWARD-HACKING INDUCTION EXPERIMENTS

To further validate the reliability of LVoG in indicating reward hacking, we conduct the opposite experiment: following the philosophy of rewarding hacking behavior and penalizing normal behavior based on the LVoG value, i.e.,

$$\lambda = \left(1 - \mathrm{LVoG}_{(0-1)}\right)^{\alpha} + 1, \qquad R_{\mathrm{fin}} = \lambda \cdot R_{\mathrm{ini}}. \tag{45}$$

This inverted modulation creates a positive feedback loop that rapidly collapses training: the model enters the hacking phase, the sample VoG becomes saliently different, the reward of the hacking samples increases, the model adopts the same hacking strategy for other samples, and is finally completely trapped in hacking. As shown in Fig. 17, the proxy reward of the induced model rises while its gold reward drops quickly, and its proxy reward even surpasses that of the unconstrained SCST baseline.

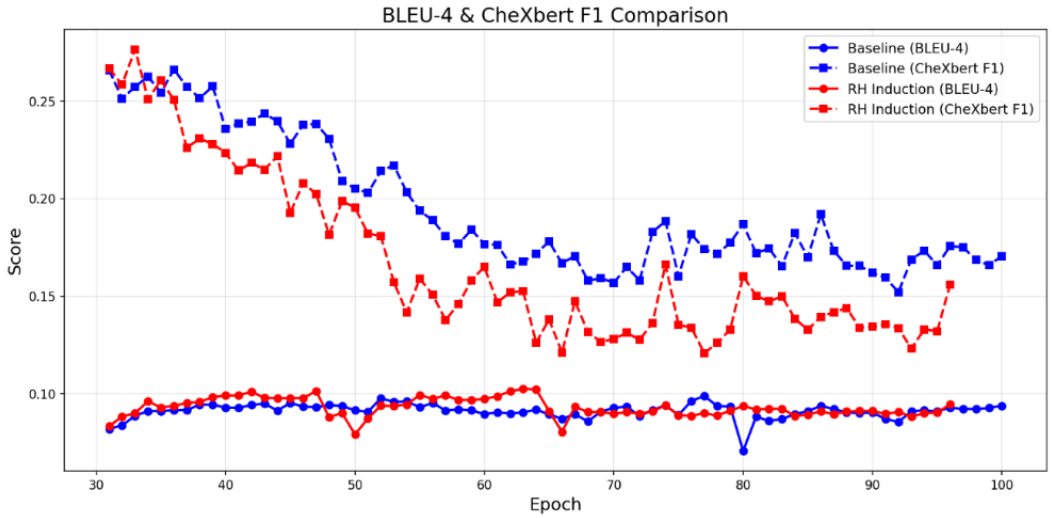

Figure 17: The proxy reward and the gold reward before and after reward-hacking induction (blue: before induction; red: after induction).

We randomly sample the predictions of part of the test set for inspection. As shown in Fig. 18, under the NLG reward (Yi et al., 2025), the model loses its ability to discriminate different samples,

completely falls into reward hacking, and collapses into a single template (yellow text). Under the RadCliQ (Ori) reward (Yu et al., 2023), the model repeats meaningless generation and even exploits the generation length to cheat the reward (red text). Optimized with our improved RadCliQ, the model still contains a few length-exploiting hacking samples but is greatly improved overall; our RHDP significantly suppresses hacking and generates healthy, clinically usable reports.

Image	Model Reward NLG（RH Induction）	Model Reward CE（RH Induction）	Model Reward RadCliQ（Ours）	Model Reward RadCliQ+RHDP
	pa and lateral views of the chest . the lungs are clear . the cardiomediastinal and hilar contours are normal . there is no pleural effusion or pneumothorax . there is no focal consolidation .	the lungs are clear without focal consolidation . no pleural effusion or pneumothorax is seen . the cardiac and mediastinal silhouettes are stable .	frontal and lateral views of the chest demonstrate notable for moderate cardiomegaly . there is mild pulmonary vascular congestion . patchy opacities in the right lung which are again seen in the right lower lung which may represent areas of atelectasis . there is no focal consolidation pleural effusion or pneumothorax . hilar and cardiomediastinal contours are within normal limits .	heart size is normal . cardiomediastinal silhouette and hilar contours are unremarkable . lung opacities likely reflect atelectasis . no focal consolidation pleural effusion or pneumothorax is present .
	pa and lateral views of the chest . the lungs are clear . the cardiomediastinal and hilar contours are normal . there is no pleural effusion or pneumothorax . there is no focal consolidation .	the cardiomediastinal silhouette and hilar contours are normal . lungs are clear without focal consolidation . there is no pleural effusion or pneumothorax .	heart size is mildly enlarged . cardiomediastinal silhouette and hilar contours are unremarkable . the lungs are clear . there is no pleural effusion focal consolidation or pneumothorax .	heart size is mildly enlarged . cardiomediastinal silhouette and hilar contours are unremarkable . lung opacities likely reflect infection contusion . small left chest wall pleural effusions are present as well as linear opacities within the right lower lung which likely represent atelectasis . no pneumothorax is identified .
	pa and lateral views of the chest . the lungs are clear . the cardiomediastinal and hilar contours are normal . there is no pleural effusion or pneumothorax . there is no focal consolidation .	heart size is mildly enlarged . the mediastinal and hilar contours are unremarkable . the pulmonary vasculature is not engorged . patchy opacities in the lung bases likely reflect subsegmental atelectasis . no focal consolidation pleural effusion or pneumothorax is present . there are no acute osseous abnormalities identified .	heart size is mildly enlarged . cardiomediastinal silhouette and hilar contours are unchanged . the right lower lobe atelectasis is noted . the right mid lung again noted likely to evaluate mild bibasilar atelectasis . there is no pleural effusion focal consolidation or pneumothorax .	heart size is mildly enlarged . cardiomediastinal silhouette and hilar contours are unremarkable . lung small bilateral pleural effusions larger on the right . moderate cardiomegaly without pulmonary edema . moderate <unk> pleural effusions are seen bilaterally . no pneumothorax is seen .
	pa and lateral views of the chest . the lungs are clear . the cardiomediastinal and hilar contours are normal . there is no pleural effusion or pneumothorax . there is no focal consolidation .	the heart size is normal . the hilar and mediastinal contours are within normal limits . there is no pneumothorax focal consolidation . there is no pleural effusion or pneumothorax . there is no pleural effusion or pneumothorax . there is no pleural effusion or pneumothorax