ADAPTIVE CLINICAL RISK ALLOCATION FOR RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinically reliable radiology report generation requires models to identify and control clinically distinct errors. Despite recent progress, existing sequence-level methods typically combine multiple clinical discrepancies with fixed weights, leaving the optimization focus unchanged as the generator's error profile evolves during training. We propose ACRA, an adaptive clinical risk allocation framework that adjusts the optimization focus as the report generator evolves. ACRA maps generated and reference reports into a shared structured finding space and tracks the generator's greedy-decoding risk profile across five reference-conditioned discrepancy categories: omission, unsupported positive finding, polarity inconsistency, laterality error, and severity error. It then applies prior-regularized online mirror ascent to allocate greater optimization mass to persistent risks while preventing excessive concentration on any single category. The resulting allocation defines a risk-aware self-critical objective for comparing sampled and greedy reports. All risk-related components are used only during training and are removed at inference. Experiments on MIMIC-CXR and IU-Xray show that ACRA improves clinical consistency and reduces reference-conditioned clinical risk while maintaining competitive report quality. These results demonstrate that adapting optimization priorities to the evolving clinical error profile enables more effective risk control without adding inference-time modules.

1 INTRODUCTION

Automatic radiology report generation translates medical images into clinically meaningful text and may reduce radiologists' reporting burden. Recent methods have improved report generation through cross-modal representation learning, memory mechanisms, disease-aware modeling, and stronger visual-language architectures (Chen et al., 2020; 2021; Wang et al., 2022; 2023; Cao et al., 2023; Wu et al., 2026). Despite this progress, clinical reliability remains challenging. A linguistically plausible report may omit an important finding, introduce an unsupported positive observation, reverse the polarity of a finding, or misstate its laterality or severity. These errors differ in clinical meaning and are not fully reflected by surface-level text similarity. Improving report quality therefore requires training objective functions that explicitly account for clinically meaningful discrepancies.

Sequence-level optimization provides a natural way to optimize report-level objectives beyond token-level maximum likelihood (Rennie et al., 2017; Miura et al., 2021; Delbrouck et al., 2022). Clinical information can be incorporated into such objectives so that generation is optimized toward report-level consistency rather than lexical agreement alone. Existing approaches, however, commonly combine multiple objectives using fixed scalar weights. This formulation assumes that the relative importance of different clinical errors remains stable throughout training. Such an assumption can be restrictive because omission, unsupported findings, polarity inconsistencies, and attribute errors differ in both prevalence and learning difficulty. Their relative difficulty may also change substantially as the generator improves.

Consider a generator that initially fails to mention many reference-supported findings. Omission may dominate its clinical discrepancy early in training and therefore deserve substantial optimization attention. As report completeness improves, omission risk may decrease rapidly, while unsupported findings or attribute inconsistencies remain difficult. Continuing to optimize these categories

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.

with the same fixed proportions can spend excessive optimization effort on risks that have already improved while under-emphasizing persistent ones. A seemingly simple alternative is to focus on whichever risk is currently largest. Such an unconstrained strategy can instead cause the objective to concentrate on a single category and lose balance across clinically distinct errors. These observations suggest that the optimization focus should evolve with the generator's clinical risk profile. Making this adaptation possible first requires clinical discrepancies to be represented as risk dimensions that can be measured and reweighted separately.

To this end, we propose **Adaptive Clinical Risk Allocation (ACRA)**, a training framework that dynamically reallocates sequence-level optimization mass across clinically distinct risks. ACRA maps generated and reference reports into a common structured finding space containing mention status, assertion state, laterality, and severity. From this representation, we define five reference-conditioned risks: omission, unsupported positive finding, polarity inconsistency, laterality error, and severity error. Each risk is normalized over its eligible opportunities using reference-conditioned eligibility. Omitted reference-positive findings are handled explicitly by the omission risk rather than being counted again as attribute errors. These risks quantify discrepancies relative to the reference report and should not be interpreted as image-grounded clinical correctness. Given this structured risk representation, ACRA adapts the allocation according to the current generator. It tracks the greedy-decoding risk profile using a smoothed online state and applies prior-regularized mirror ascent to obtain a distribution over risk categories. Risks that remain persistent receive greater optimization mass, while regularization toward a reference prior prevents excessive concentration on a single category. The resulting allocation is shared when scoring sampled and greedy reports in self-critical sequence training, allowing the training objective to change as the generator evolves. ACRA therefore modifies the optimization process without changing the underlying report generator. All structured extraction and risk allocation components are used only during training, leaving inference unchanged. Experiments on MIMIC-CXR and IU-Xray show that ACRA improves clinical consistency while maintaining competitive generation quality. Our main contributions are:

- We formulate clinical risk control in sequence-level radiology report optimization as an adaptive allocation problem, where the optimization focus evolves with the generator's changing error profile.
- We introduce ACRA, which decomposes reference-conditioned discrepancies into five structured clinical risks and dynamically allocates optimization mass through prior-regularized online mirror ascent, without adding inference-time modules.
- Experiments on MIMIC-CXR and IU-Xray demonstrate lower reference-conditioned risk and improved clinical consistency, supported by controlled comparisons, ablations, and training-dynamics analysis.

2 Related Work

Radiology Report Generation. Radiology report generation translates medical images into clinically meaningful free-text reports. Early neural approaches mainly improved visual-text alignment and long-range generation. R2Gen introduced a memory-driven Transformer, while R2GenCMN modeled cross-modal interactions through shared memory representations (Chen et al., 2020; 2021). Subsequent methods introduced more structured visual and semantic modeling. XProNet learns cross-modal prototypes, while METransformer and MMTN improve multimodal interaction through expert tokens and memory-based representations (Wang et al., 2022; 2023; Cao et al., 2023). Recent work further incorporates memory alignment, clinical concepts, and disease-aware modeling to improve report accuracy (Huang et al., 2023; Shen et al., 2024; Wu et al., 2026). These methods primarily improve the generator or its visual-semantic representations. ACRA instead addresses the optimization process, studying how training emphasis among clinically distinct errors should evolve without modifying the inference architecture.

Clinical Consistency and Structured Evaluation. Conventional language metrics provide limited information about clinical correctness, since lexically similar reports can differ in clinically important details. CheXpert and CheXbert represent reports through observation labels and assertion states (Irvin et al., 2019; Smit et al., 2020), while RadGraph extracts clinical entities and relations and Chest ImaGenome provides anatomy-aware annotations (Jain et al., 2021; Wu et al., 2021). More recent metrics, including RadCliQ, GREEN, FineRadScore, and RaTEScore, aim to better

reflect clinically meaningful report quality or provide finer-grained error analysis (Yu et al., 2023; Ostmeier et al., 2024; Huang et al., 2024; Zhao et al., 2024). These approaches improve the characterization of clinical disagreement, but do not determine how distinct errors should be weighted during optimization. ACRA uses structured discrepancies for this training-time purpose rather than introducing another general-purpose evaluation metric.

Sequence-Level and Risk-Aware Optimization. Maximum-likelihood training optimizes token-level prediction and does not directly target report-level quality. Self-critical sequence training instead optimizes sequence-level rewards using greedy decoding as a baseline (Rennie et al., 2017), and has been applied to radiology reporting with clinical, factual, and semantic objectives (Miura et al., 2021; Delbrouck et al., 2022). Such objectives better align training with properties that token likelihood or lexical metrics may miss. A remaining issue is how multiple clinical objectives are combined. Fixed coefficients assume that their relative importance remains stable, although different error categories may improve at different rates. ACRA therefore treats risk allocation as an online training variable, adapting the optimization emphasis to the current generator while regularizing against excessive concentration on a single risk. This allocation is used only during sequence-level training and adds no inference-time components.

3 Approach

3.1 Overview

Given a chest radiograph I, an autoregressive generator G_θ defines a distribution $p_\theta(Y \mid I)$ over reports $Y = [y_1, \ldots, y_T]$. Our framework is agnostic to the specific generator architecture and does not modify its decoding structure. At training iteration t, G_θ produces a sampled report Y^s by multinomial sampling and a greedy report Y^g by token-wise argmax decoding. Together with the reference Y^*, both reports are mapped to a common structured clinical space, where discrepancies are represented as category-specific risks.

ACRA uses Y^g to estimate the current greedy-decoding risk profile and maintains an adaptive distribution $\boldsymbol{\alpha}^{(t)}$ over risk categories. Persistent risks receive greater optimization mass, while regularization toward a reference prior limits excessive concentration. The resulting allocation is fixed within the current generator update and shared when scoring Y^s and Y^g. Their score difference defines the self-critical advantage used to update θ. Sharing the same allocation is important because the advantage should compare two reports under the same clinical objective. The overall procedure therefore changes the optimization process while leaving inference unchanged.

3.2 Structured Clinical Risk Representation

Structured finding states. We use the $K = 14$ CheXbert observation categories as the finding inventory $\mathcal{F} = \{f_k\}_{k=1}^K$. For a report Y, the clinical extractor returns

$$E(Y) = \{e_k(Y)\}_{k=1}^K, \qquad e_k(Y) = (m_k, z_k, \ell_k, s_k), \tag{1}$$

where $m_k \in \{0, 1\}$ denotes mention status, $z_k \in \{\mathrm{pos}, \mathrm{neg}, \mathrm{unc}\}$ is the assertion state when mentioned, and ℓ_k and s_k denote laterality and severity. Unavailable or inapplicable attributes are assigned $\emptyset$.

We obtain these states with a fixed hybrid pipeline. RadGraph extracts observation entities, assertions, and anatomy relations Jain et al. (2021); entities are normalized to $\mathcal{F}$ with a synonym dictionary, and CheXbert provides a report-level fallback when no valid RadGraph match is available Smit et al. (2020). Laterality and severity are extracted from associated sentence and anatomy spans using deterministic lexicons, with attribute normalization following Chest ImaGenome Wu et al. (2021). Multiple mentions mapped to the same finding are consolidated deterministically, yielding one state per finding. The same pipeline is applied to Y^s, Y^g, and Y^*. The *No Finding* category is excluded from positive-finding matching and category-risk computation; detailed aggregation and conflict-resolution rules are given in the implementation details.

Reference-conditioned clinical risk. We model five discrepancy categories $\mathcal{C} = \{\mathrm{omi}, \mathrm{unsup}, \mathrm{pol}, \mathrm{lat}, \mathrm{sev}\}$, corresponding to omission, unsupported positive finding, polarity

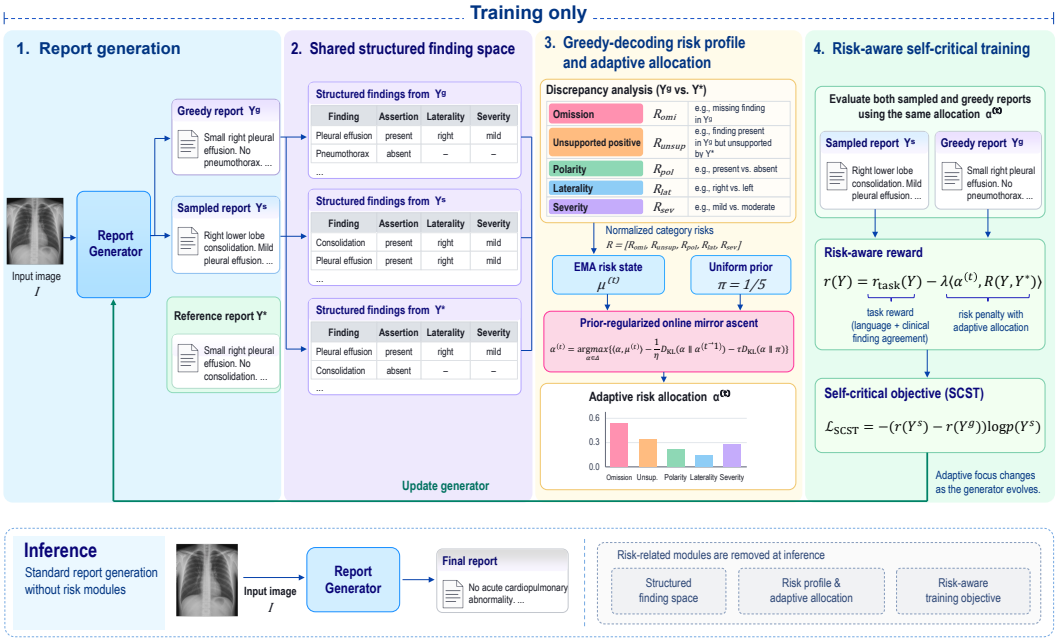


Figure 1: Overview of ACRA. A sampled report, a greedy report, and the reference are mapped to a common structured finding space. Greedy-decoding risks update a smoothed risk state and a prior-regularized allocation over five discrepancy categories. The same allocation scores sampled and greedy reports, whose difference provides the self-critical advantage. Only the report generator is retained at inference.

Table 1: Eligibility and error indicators for the five reference-conditioned clinical risks.

Category	Eligibility a_c^k	Error indicator δ_c^k
omi	χ_k^*	$a_{\text{omi}}^k \mathbf{1}[m_k = 0]$
unsup	$\mathbf{1}[m_k^* = 0]$	$a_{\text{unsup}}^k \chi_k$
pol	m_k^*	$a_{\text{pol}}^k m_k \mathbf{1}[z_k \neq z_k^*]$
lat	$\chi_k^* \mathbf{1}[\ell_k^* \neq \emptyset]$	$a_{\text{lat}}^k \chi_k \mathbf{1}[\ell_k = \emptyset \vee \ell_k \neq \ell_k^*]$
sev	$\chi_k^* \mathbf{1}[s_k^* \neq \emptyset]$	$a_{\text{sev}}^k \chi_k \mathbf{1}[s_k = \emptyset \vee s_k \neq s_k^*]$

inconsistency, laterality error, and severity error. These quantities measure disagreement with the reference, not image-grounded clinical correctness. In particular, unsup is a reference-based proxy for hallucination. Uncertainty contributes to polarity inconsistency when a finding is mentioned in both reports, but uncertainty-specific omission and unsupported errors are not modeled separately. Let $\chi_k = \mathbf{1}[m_k = 1 \wedge z_k = \text{pos}]$, with superscript $*$ denoting reference states. For each category, Table 1 specifies an eligibility indicator a_c^k and an error indicator δ_c^k. Eligibility is primarily reference-conditioned, with omitted reference-positive findings handled by the omission risk rather than attribute risks. Omission and unsupported-positive risks operate at the finding level, polarity compares assertion states for mentioned findings, and laterality and severity are evaluated only when the reference supplies the corresponding attribute. A missing laterality or severity value is counted as an attribute error only when the generated report positively mentions a finding for which the reference specifies that attribute; if the finding itself is absent, the discrepancy is counted as an omission. With uniform finding weights $\omega_k = 1$, category risk is normalized by its eligible mass:

$$D_c(Y^*) = \sum_{k=1}^{K} \omega_k a_c^k, \qquad \bar{R}_c(Y, Y^*) = \begin{cases} \dfrac{\sum_k \omega_k \delta_c^k}{D_c(Y^*)}, & D_c(Y^*) > 0, \\ 0, & D_c(Y^*) = 0. \end{cases} \tag{2}$$

Thus $\bar{R}_c \in [0, 1]$, and the report-level risk vector is $\mathbf{R}(Y, Y^*) = [\bar{R}_c(Y, Y^*)]_{c \in \mathcal{C}}$.

3.3 Online Minimax Risk Calibration

Fixed risk weights assume that the relative importance of clinical errors remains stable during training. We instead view risk allocation as an online minimax problem in which a risk player emphasizes categories that remain difficult for the current generator. The motivating regularized objective is

$$\min_{\theta} \max_{\boldsymbol{\alpha} \in \Delta_{|\mathcal{C}|}} \mathbb{E}[-r_{\text{task}}(Y, Y^*) + \lambda \langle \boldsymbol{\alpha}, \mathbf{R}(Y, Y^*) \rangle] - \lambda \tau D_{\text{KL}}(\boldsymbol{\alpha} \| \boldsymbol{\pi}), \tag{3}$$

where the expectation is over training image-reference pairs and generated reports, $\boldsymbol{\pi}$ is a positive reference distribution, and $\tau \geq 0$ controls regularization. We use the uniform prior $\pi_c = 1/|\mathcal{C}|$. Rather than solving the inner maximization to convergence at every generator step, we track the risk player online from the current greedy-decoding profile. Greedy decoding provides a deterministic, low-variance proxy for the generator's current behavior and naturally matches the self-critical baseline.

For a risk profile $\boldsymbol{\mu}$, the risk player maximizes $\langle \boldsymbol{\alpha}, \boldsymbol{\mu} \rangle - \tau D_{\text{KL}}(\boldsymbol{\alpha} \| \boldsymbol{\pi})$. Directly averaging report-level risks would treat non-evaluable categories as zeros. We therefore aggregate errors over eligible opportunities across a mini-batch of B greedy-reference pairs:

$$b_c^{(t)} = \frac{\sum_{i=1}^{B} \sum_{k=1}^{K} \omega_k \delta_{c,i}^k}{\sum_{i=1}^{B} \sum_{k=1}^{K} \omega_k a_{c,i}^k}, \qquad D_c^{(t)} > 0, \tag{4}$$

where $D_c^{(t)}$ denotes the denominator. We smooth this estimate as

$$\mu_c^{(t)} = \begin{cases} \rho \mu_c^{(t-1)} + (1-\rho) b_c^{(t)}, & D_c^{(t)} > 0, \\ \mu_c^{(t-1)}, & D_c^{(t)} = 0, \end{cases} \tag{5}$$

with $\boldsymbol{\mu}^{(0)} = \mathbf{0}$ and $\rho \in [0, 1)$. Starting from $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\pi}$, we perform one prior-regularized KL mirror-ascent step:

$$\boldsymbol{\alpha}^{(t)} = \arg \max_{\boldsymbol{\alpha} \in \Delta_{|\mathcal{C}|}} \left\{ \eta \langle \boldsymbol{\alpha}, \boldsymbol{\mu}^{(t)} \rangle - D_{\text{KL}}(\boldsymbol{\alpha} \| \boldsymbol{\alpha}^{(t-1)}) - \eta \tau D_{\text{KL}}(\boldsymbol{\alpha} \| \boldsymbol{\pi}) \right\}. \tag{6}$$

Its closed-form solution is

$$\alpha_c^{(t)} = \frac{\left[\alpha_c^{(t-1)} \exp(\eta \mu_c^{(t)}) \pi_c^{\eta\tau} \right]^{1/(1+\eta\tau)}}{\sum_{j \in \mathcal{C}} \left[\alpha_j^{(t-1)} \exp(\eta \mu_j^{(t)}) \pi_j^{\eta\tau} \right]^{1/(1+\eta\tau)}}, \tag{7}$$

where $\eta > 0$ is the ascent step size. The proximal KL term limits abrupt movement from the previous allocation, while the prior KL term discourages concentration on a single risk. Hence high persistent risks receive more mass without allowing the allocation to collapse to a single category. When $\tau = 0$, the update reduces to standard exponentiated gradient. For the subsequent generator update, $\boldsymbol{\alpha}^{(t)}$ is stop-gradient and shared by the sampled and greedy reports. Their calibrated risk is

$$Q_{\text{risk}}^{(t)}(Y, Y^*) = \left\langle \text{sg}[\boldsymbol{\alpha}^{(t)}], \mathbf{R}(Y, Y^*) \right\rangle \in [0, 1]. \tag{8}$$

3.4 Risk-Aware Sequence Objective

The sequence objective combines conventional report utility with the adaptive structured risk. Language utility is $r_{\text{lang}} = \frac{1}{2}(r_{\text{B4}} + r_{\text{ROUGE-L}})$, with both metrics scaled to $[0, 1]$ Papineni et al. (2002); Lin (2004). For positive finding agreement, define $\mathcal{P}(Y) = \{f_k : m_k = 1, z_k = \text{pos}\}$. We use

$$F_{\text{clin}}(Y, Y^*) = \begin{cases} 1, & |\mathcal{P}(Y)| = |\mathcal{P}(Y^*)| = 0, \\ \dfrac{2|\mathcal{P}(Y) \cap \mathcal{P}(Y^*)|}{|\mathcal{P}(Y)| + |\mathcal{P}(Y^*)|}, & \text{otherwise}, \end{cases} \tag{9}$$

which is zero when only one report contains positive findings. We then define

$$r_{\text{task}}(Y, Y^*) = \gamma r_{\text{lang}}(Y, Y^*) + (1-\gamma) F_{\text{clin}}(Y, Y^*), \tag{10}$$

$$S^{(t)}(Y, Y^*) = r_{\text{task}}(Y, Y^*) - \lambda Q_{\text{risk}}^{(t)}(Y, Y^*), \tag{11}$$

where $\gamma \in [0, 1]$ and $\lambda \geq 0$. F_{clin} provides a coarse measure of overall positive-finding agreement, while $Q_{\text{risk}}^{(t)}$ explicitly separates clinically different errors. The coefficient γ remains fixed, whereas $\boldsymbol{\alpha}^{(t)}$ adapts the relative contribution of the five risk categories as training progresses.

3.5 Self-Critical Sequence Optimization

Because $S^{(t)}$ depends on discrete text and nondifferentiable clinical extraction, we optimize it with self-critical sequence training Rennie et al. (2017). Using the same $\boldsymbol{\alpha}^{(t)}$ for Y^s and Y^g, we compute

$$A^{(t)} = S^{(t)}(Y^s, Y^*) - S^{(t)}(Y^g, Y^*), \tag{12}$$

$$\ell_\theta(Y^s \mid I) = \frac{1}{T_s} \sum_{j=1}^{T_s} \log p_\theta(y_j^s \mid y_{<j}^s, I), \tag{13}$$

$$\mathcal{L}_{\text{seq}} = -\text{sg}[A^{(t)}]\, \ell_\theta(Y^s \mid I), \tag{14}$$

where length normalization prevents the likelihood magnitude from scaling directly with report length. This normalization is an optimization choice and does not alter the sequence-level reward.

Training uses two stages. We first optimize the generator with standard negative log-likelihood $\mathcal{L}_{\text{NLL}}$, then fine-tune it with

$$\mathcal{L} = \beta \mathcal{L}_{\text{NLL}} + (1 - \beta)\mathcal{L}_{\text{seq}}, \tag{15}$$

where $\beta \in [0, 1]$. Within each sequence-training iteration, the generator first produces Y^s and Y^g; the greedy report updates $\boldsymbol{\mu}^{(t)}$ and then $\boldsymbol{\alpha}^{(t)}$; the fixed allocation is finally used to compute $A^{(t)}$ and update θ. At inference, the reference report, extractor, risk state, allocation, and self-critical baseline are removed, leaving only the underlying report generator.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate ACRA on two widely used radiology report generation benchmarks, MIMIC-CXR Johnson et al. (2019) and IU-Xray Demner-Fushman et al. (2016). MIMIC-CXR contains 377,110 chest X-ray images from 227,835 studies and uses the official splits. IU-Xray contains 7,470 images and 3,955 reports, for which we follow the standard 7:1:2 train/validation/test split. We report BLEU-1 to BLEU-4 Papineni et al. (2002), METEOR Banerjee & Lavie (2005), and ROUGE-L Lin (2004) for conventional generation quality. Clinical consistency is evaluated with CheXbert micro-F1 Smit et al. (2020), RadGraph-F1 Jain et al. (2021), positive-finding F_{clin}, and the five reference-conditioned risks.

Implementation. We instantiate G_θ with R2GenCMN Chen et al. (2021) using an ImageNet-pretrained ResNet-101 encoder. Adam uses learning rates of 10^{-6} for the visual encoder and 10^{-5} for the remaining parameters. We set $\gamma = 0.8$, $\lambda = 0.5$, $\rho = 0.9$, $\eta = 0.1$, $\tau = 0.5$, and $\beta = 0.01$, with uniform finding weights and a uniform risk prior. Sequence training uses one sampled and one greedy report per image. At inference, all risk-related modules are removed and reports are generated with beam size 3. Experiments are conducted on NVIDIA A6000 GPUs. Additional details are provided in Appendix A.1.

4.2 Comparison with State-of-the-Art Methods

We compare ACRA with representative RRG methods spanning memory-based, cross-modal, pre-training, and disease-aware paradigms. Published results are taken from the corresponding papers. To isolate the effect of our training objective from architectural differences, we additionally reproduce R2GenCMN using the same backbone, preprocessing, NLL pretraining, and decoding protocol as ACRA. Tables 2 and 3 show that ACRA remains competitive with recent architecture- and pre-training-oriented methods while retaining the original R2GenCMN inference architecture. On IU-Xray, ACRA achieves the best performance on all six generation metrics. On MIMIC-CXR, it ranks first on BLEU-2, BLEU-3, BLEU-4, and ROUGE-L, and second on BLEU-1 and METEOR. More importantly, ACRA consistently improves over R2GenCMN (reprod.) on every reported metric across both datasets. Since the two models share the same generator architecture and inference protocol, these gains can be attributed to the proposed risk-aware sequence-level optimization. The results also indicate that introducing explicit clinical risk control does not compromise conventional report generation quality.

Table 2: Comparison on IU-Xray using standard generation metrics. B-n, MTR, and R-L denote BLEU-n, METEOR, and ROUGE-L. Higher is better. Best results are bold and second-best results are underlined.

Method	Venue	B-1	B-2	B-3	B-4	MTR	R-L
R2Gen Chen et al. (2020)	EMNLP'20	0.470	0.304	0.219	0.165	0.187	0.371
R2GenCMN Chen et al. (2021)	ACL'21	0.475	0.309	0.222	0.170	0.191	0.375
XProNet Wang et al. (2022)	ECCV'22	<u>0.525</u>	<u>0.357</u>	<u>0.262</u>	<u>0.199</u>	0.220	<u>0.411</u>
METransformer Wang et al. (2023)	CVPR'23	0.483	0.322	0.228	0.172	0.192	0.380
MMTN Cao et al. (2023)	AAAI'23	0.486	0.321	0.232	0.175	–	0.375
MAN Shen et al. (2024)	AAAI'24	0.501	0.328	0.230	0.170	0.213	0.386
MambaXray-VL-Large Wang et al. (2025)	CVPR'25	0.491	0.330	0.241	0.185	0.216	0.371
Disease-Aware Dual-Stage Wu et al. (2026)	AAAI'26	0.495	0.328	0.242	0.187	–	0.384
R2GenCMN (reprod.)	–	0.515	<u>0.357</u>	0.258	0.193	<u>0.223</u>	<u>0.411</u>
ACRA	Ours	**0.535**	**0.372**	**0.269**	**0.202**	**0.225**	**0.413**

Table 3: Comparison on MIMIC-CXR using standard generation metrics. B-n, MTR, and R-L denote BLEU-n, METEOR, and ROUGE-L. Higher is better. Best results are bold and second-best results are underlined.

Method	Venue	B-1	B-2	B-3	B-4	MTR	R-L
R2Gen Chen et al. (2020)	EMNLP'20	0.353	0.218	0.145	0.103	0.142	0.277
R2GenCMN Chen et al. (2021)	ACL'21	0.353	0.218	0.148	0.106	0.142	0.278
XProNet Wang et al. (2022)	ECCV'22	0.344	0.215	0.146	0.105	0.138	0.279
METransformer Wang et al. (2023)	CVPR'23	0.386	0.250	0.169	0.124	0.152	0.291
MMTN Cao et al. (2023)	AAAI'23	0.379	0.238	0.159	0.116	0.161	0.283
MAN Shen et al. (2024)	AAAI'24	0.396	0.244	0.162	0.115	0.151	0.274
MambaXray-VL-Large Wang et al. (2025)	CVPR'25	0.422	0.268	0.184	<u>0.133</u>	**0.167**	0.289
Disease-Aware Dual-Stage Wu et al. (2026)	AAAI'26	**0.428**	<u>0.272</u>	<u>0.187</u>	0.131	–	0.291
R2GenCMN (reprod.)	–	0.407	0.258	0.177	0.127	0.155	<u>0.296</u>
ACRA	Ours	<u>0.426</u>	**0.278**	**0.188**	**0.134**	<u>0.164</u>	**0.303**

4.3 Clinical Risk Analysis

We next examine whether the improvements in generation quality are accompanied by better clinical consistency, and isolate the contributions of sequence-level training, explicit risk modeling, and adaptive allocation. All variants use the same R2GenCMN backbone. *R2GenCMN (reprod.)* uses NLL training; *SCST task-only* optimizes r_{task} without explicit clinical risks; *Fixed risk* uses the same five risk categories as ACRA with $\alpha^{(t)} = \pi$; and *ACRA* applies the full adaptive allocation mechanism. For dataset-level evaluation, each category is aggregated over all eligible test-set opportunities. We further report an allocation-independent macro average that assigns equal evaluation weight to the five categories:

$$\widehat{R}_c = \frac{\sum_{i,k} \omega_k \delta_{c,i}^k}{\sum_{i,k} \omega_k a_{c,i}^k}, \qquad R_{\text{avg}} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \widehat{R}_c. \tag{16}$$

Table 4 shows a consistent progression across the four training configurations on both datasets. Task-only SCST improves all three clinical consistency metrics, but reduces R_{avg} by only about 4% relative to NLL training. Introducing explicit structured risks produces substantially larger gains: Fixed risk reduces R_{avg} by 11.7% on IU-Xray and 11.6% on MIMIC-CXR relative to task-only SCST. This indicates that explicitly modeling clinically distinct discrepancies provides benefits beyond generic sequence-level optimization. ACRA further reduces R_{avg} by 10.5% on IU-Xray and 10.6% on MIMIC-CXR relative to Fixed risk, while consistently improving CheXbert-F1, RadGraph-F1, and F_{clin}. Since the two variants share the same generator, risk definitions, and sequence objective, this comparison isolates the contribution of adaptive risk allocation under otherwise matched settings.

Table 4: Clinical consistency and reference-conditioned risk on IU-Xray and MIMIC-CXR. CheX-F1 and RG-F1 denote CheXbert micro-F1 and RadGraph-F1. Higher is better for the first three metrics and lower is better for risks.

Dataset	Method	CheX-F1	RG-F1	F_{clin}	Avg.	Omi.	Unsup.	Pol.	Lat.	Sev.
IU-Xray	R2GenCMN (reprod.)	0.425	0.198	0.378	0.214	0.286	0.248	0.169	0.116	0.252
	SCST task-only	0.432	0.202	0.389	0.205	0.278	0.238	0.162	0.113	0.234
	Fixed risk	0.445	0.211	0.405	0.181	0.247	0.208	0.145	0.101	0.204
	ACRA	**0.454**	**0.218**	**0.419**	**0.162**	**0.218**	**0.181**	**0.132**	**0.093**	**0.186**
MIMIC-CXR	R2GenCMN (reprod.)	0.367	0.191	0.334	0.233	0.305	0.276	0.183	0.127	0.274
	SCST task-only	0.375	0.195	0.346	0.224	0.297	0.265	0.177	0.124	0.257
	Fixed risk	0.389	0.204	0.365	0.198	0.269	0.232	0.157	0.111	0.221
	ACRA	**0.399**	**0.211**	**0.381**	**0.177**	**0.238**	**0.204**	**0.144**	**0.102**	**0.197**

Figure 2 further reveals that the five discrepancy categories exhibit different levels of difficulty on both datasets. Omission remains the largest risk, whereas laterality is consistently lower. Task-only SCST changes the overall profile only modestly, Fixed risk produces broader reductions, and ACRA further improves every category. Notably, the gain from adaptive allocation is not confined to the dominant risks. Omission and unsupported-positive errors decrease together with polarity, laterality, and severity errors. Together with the consistent gains in CheXbert-F1 and RadGraph-F1, these results suggest that ACRA improves the overall clinical consistency profile rather than optimizing a single risk category at the expense of the others.

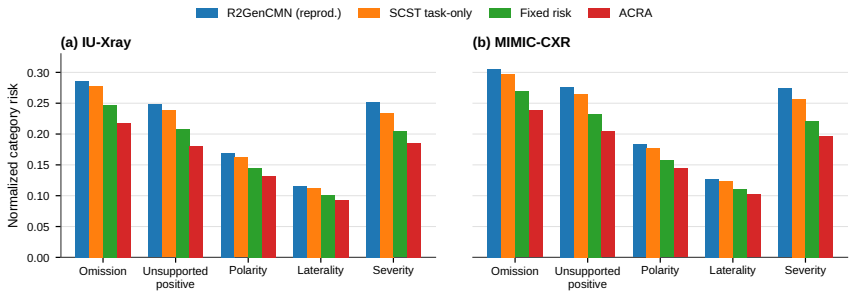


Figure 2: Reference-conditioned category risks on IU-Xray and MIMIC-CXR. Lower values indicate lower risk.

4.4 Ablation Study

We evaluate the contributions of the three components in adaptive risk allocation on MIMIC-CXR. *w/o Adaptive Allocation* fixes $\alpha^{(t)} = \pi$; *w/o Prior Regularization* sets $\tau = 0$; and *w/o Risk Smoothing* replaces the EMA state with the current mini batch risk estimate.

Table 5: Ablation of ACRA on MIMIC-CXR. Higher is better for clinical consistency metrics and lower is better for risks. Best values are shown in bold.

Variant	CheX-F1	RG-F1	F_{clin}	Avg.	Omi.	Unsup.	Pol.	Lat.	Sev.
w/o Adaptive Allocation	0.389	0.204	0.365	0.198	0.269	0.232	0.157	0.111	0.221
w/o Prior Regularization	0.394	0.207	0.373	0.185	**0.231**	**0.198**	0.153	0.113	0.230
w/o Risk Smoothing	0.396	0.209	0.376	0.183	0.244	0.211	0.149	0.108	0.203
ACRA	**0.399**	**0.211**	**0.381**	**0.177**	0.238	0.204	**0.144**	**0.102**	**0.197**

Table 5 shows that adaptive allocation provides the largest overall contribution. Fixing the allocation increases R_{avg} from 0.177 to 0.198 and degrades all three clinical consistency metrics, confirming that the gain of ACRA cannot be explained by structured risk modeling alone. Removing prior

regularization yields slightly lower omission and unsupported positive risks, but increases polarity, laterality, and severity risks, resulting in a higher R_{avg} of 0.185. This tradeoff is consistent with the prior discouraging excessive concentration on dominant risk categories. Removing risk smoothing also degrades both aggregate risk and clinical consistency, indicating that the smoothed risk state provides a more reliable signal for adaptive allocation. Overall, the full model achieves the lowest average risk and the best clinical consistency across all three metrics.

4.5 Adaptive Risk Allocation Analysis

We further examine how the estimated clinical risks and their corresponding allocation evolve during sequence-level training. Figure 3 shows the smoothed greedy-decoding risk state $\boldsymbol{\mu}^{(t)}$ and the adaptive allocation $\boldsymbol{\alpha}^{(t)}$ on MIMIC-CXR.

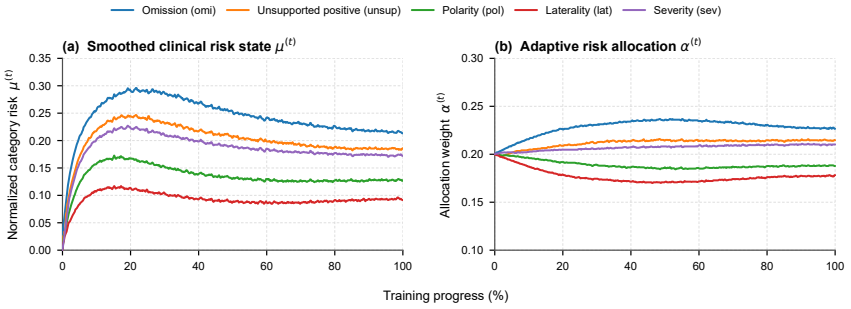


Figure 3: Training dynamics of the smoothed clinical risk state $\boldsymbol{\mu}^{(t)}$ (left) and adaptive risk allocation $\boldsymbol{\alpha}^{(t)}$ (right) on MIMIC-CXR. All allocations start from the uniform prior $\alpha_c^{(0)} = 0.2$.

As shown in Figure 3(a), the smoothed risk state first approaches the generator's initial greedy-decoding profile and then declines as sequence-level training progresses. The five categories exhibit distinct trajectories: omission remains the largest risk, unsupported-positive and severity risks occupy an intermediate range, while polarity and laterality remain lower. Their relative levels also change over training, providing empirical evidence that a fixed risk allocation cannot fully reflect the evolving clinical error profile. Figure 3(b) shows that the learned allocation responds to these differences. Starting from the uniform prior, more mass is assigned to persistent high-risk categories, particularly omission, whereas lower-risk categories receive smaller weights. The allocation changes smoothly over training and remains distributed across all five risks rather than concentrating on a single category. Together with the ablation results in Table 5, these dynamics support the intended roles of adaptive allocation, temporal smoothing, and prior regularization.

5 Conclusion

We presented ACRA, an adaptive clinical risk allocation framework for radiology report generation. ACRA represents generated and reference reports in a structured finding space and dynamically reallocates optimization emphasis according to the evolving clinical risk profile of the generator through prior-regularized online mirror ascent. The resulting allocation is incorporated into self-critical sequence training, while all risk-related components are removed at inference. Experiments on MIMIC-CXR and IU-Xray show that ACRA improves clinical consistency and reduces reference-conditioned clinical risk while maintaining competitive generation quality. Controlled comparisons and ablations confirm the benefits of structured risk modeling, adaptive allocation, prior regularization, and risk smoothing. Training dynamics further show that different clinical risks evolve at different rates and that the learned allocation responds accordingly. These results support adaptive risk allocation as an effective alternative to fixed clinical risk scalarization for sequence-level radiology report optimization.

6 AI USE STATEMENT

We used ChatGPT to assist with language polishing and stylistic refinement of the manuscript. All AI-assisted revisions were manually reviewed and edited by the authors to ensure that the technical content, terminology, claims, and presentation accurately reflect the work conducted in this study. The authors take full responsibility for the final content of this paper, including any text produced or revised with the assistance of generative AI.

7 ETHICS STATEMENT

This study uses existing publicly available radiology datasets and does not involve new patient recruitment, intervention, or data collection. Our work is intended to improve the clinical consistency of automatic radiology report generation, but generated reports may still contain omissions, unsupported findings, or other clinically relevant errors. Moreover, the clinical risks considered in ACRA are defined relative to reference reports and should not be interpreted as direct image-grounded clinical adjudication or as a substitute for expert assessment. We therefore do not claim that the proposed system is suitable for autonomous clinical deployment. Any practical use of automatic report generation should involve appropriate validation and qualified clinical oversight. As with other data-driven medical AI systems, performance may also depend on the characteristics of the training data and may not transfer uniformly across institutions, imaging protocols, or patient populations.

8 REPRODUCIBILITY STATEMENT

We provide the information required to reproduce ACRA throughout the main paper and appendix. Section 3 describes the complete optimization framework, including the structured clinical risk representation, online risk calibration, prior-regularized mirror-ascent update, and risk-aware self-critical training objective. Dataset construction, evaluation protocols, and the main training and inference settings are described in Section 4.1. Additional implementation details, including optimizer settings, batch sizes, risk-calibration hyperparameters, finding-state aggregation rules, and sequence-level training configuration, are reported in Appendix A.1. We use publicly available MIMIC-CXR and IU-Xray datasets and follow the data splits specified in the experimental setup. Together, these details specify the model configuration, training procedure, clinical risk computation, and evaluation protocol used in our experiments.

REFERENCES

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. Association for Computational Linguistics, 2005.

Yiming Cao, Lizhen Cui, Lei Zhang, Fuqiang Yu, Zhen Li, and Yonghui Xu. MMTN: Multi-modal memory transformer network for image-report consistent medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 277–285, 2023. doi: 10.1609/aaai.v37i1.25100.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1439–1449. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.112.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 5904–5914. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.459.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis P. Langlotz. Improving the factual correctness of radiology report generation with semantic

rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360. Association for Computational Linguistics, 2022.

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.

Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. FineRad-Score: A radiology report line-by-line evaluation technique generating corrections with severity scores. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v252/huang24a.html.

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*, pp. 19809–19818, 2023.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019. doi: 10.1609/aaai.v33i01.3301590.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

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. doi: 10.1038/s41597-019-0322-0.

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

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5288–5304. Association for Computational Linguistics, 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*, pp. 374–390, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

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*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7008–7024, 2017. doi: 10.1109/CVPR.2017.131.

Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4776–4783, 2024. doi: 10.1609/aaai.v38i5.28279.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. 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*, pp. 1500–1519. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.117.

Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report generation. In *Proceedings of the European Conference on Computer Vision*, pp. 563–579. Springer, 2022. doi: 10.1007/978-3-031-19833-5_33.

Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, and Jin Tang. CXPMRG-Bench: Pre-training and benchmarking for x-ray medical report generation on chexpert plus dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5123–5133, 2025.

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*, pp. 11558–11567, 2023.

Joy T. Wu, Nkechinyere N. Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, Jasper S. Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo A. Celi, and Mehdi Moradi. Chest imagenome dataset for clinical reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Puzhen Wu, Hexin Dong, Yi Lin, Yihao Ding, and Yifan Peng. A disease-aware dual-stage framework for chest x-ray report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33953–33961, 2026. doi: 10.1609/aaai.v40i40.40688.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023. doi: 10.1016/j.patter.2023.100802. URL https://doi.org/10.1016/j.patter.2023.100802.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15004–15019, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.836. URL https://aclanthology.org/2024.emnlp-main.836/.

A Additional Details and Results

A.1 Implementation Details

We follow the standard R2GenCMN configuration Chen et al. (2021) and use an ImageNet-pretrained ResNet-101 as the visual encoder. The model is optimized with Adam using a learning rate of 1×10^{-6} for the visual encoder and 1×10^{-5} for the remaining parameters, a weight decay of 5×10^{-5}, and Adam coefficients $(0.9, 0.98)$. The batch sizes are 6 for MIMIC-CXR and 8 for IU-Xray.

Training consists of two stages. We first optimize the report generator with the standard negative log-likelihood objective $\mathcal{L}_{\mathrm{NLL}}$. The resulting model is then fine-tuned with the combined sequence-level objective

$$\mathcal{L} = \beta \mathcal{L}_{\mathrm{NLL}} + (1 - \beta)\mathcal{L}_{\mathrm{seq}}, \tag{17}$$

where $\beta = 0.01$. At each sequence-level update, one report is generated by multinomial sampling and one report is produced by greedy decoding. The greedy report is used to update the online

clinical risk state and adaptive allocation, while the same allocation is used to score both reports in the self-critical advantage.

For the task reward, we set $\gamma = 0.8$ and $\lambda = 0.5$. The online risk state uses an EMA coefficient of $\rho = 0.9$, and the prior-regularized mirror-ascent update uses step size $\eta = 0.1$ and regularization strength $\tau = 0.5$. Finding weights are uniform, $\omega_k = 1$, and the allocation prior is also uniform,

$$\pi_c = \frac{1}{|\mathcal{C}|}. \tag{18}$$

Accordingly, the initial allocation assigns equal mass to the five risk categories. BLEU-4 and ROUGE-L rewards are represented on the $[0, 1]$ scale when constructing the language component of the sequence reward.

All structured clinical extraction and risk-calibration components are used only during training. At inference, the reference report, clinical extractor, risk state, adaptive allocation, and self-critical baseline are removed, leaving the original report generator unchanged. Reports are generated using beam search with beam size 3. All experiments are conducted on NVIDIA A6000 GPUs with 48 GB memory.

A.2 Structured Finding Extraction and Normalization

ACRA computes its clinical risks in a shared structured finding space rather than directly from free-text reports. We use the $K = 14$ CheXbert observation categories as the common finding inventory $\mathcal{F} = \{f_k\}_{k=1}^{K}$. For each finding f_k in a report Y, the extractor produces

$$e_k(Y) = (m_k, z_k, \ell_k, s_k), \tag{19}$$

where m_k denotes whether the finding is mentioned, $z_k \in \{\text{pos}, \text{neg}, \text{unc}\}$ denotes its assertion state, and ℓ_k and s_k denote laterality and severity, respectively. Unavailable or inapplicable attributes are represented by $\emptyset$.

Finding and assertion extraction. RadGraph Jain et al. (2021) is first applied to each report to identify observation entities, assertion states, and anatomy relations. Extracted observation entities are normalized to the common finding inventory through a fixed synonym dictionary. When no valid RadGraph entity can be mapped to a finding category, CheXbert Smit et al. (2020) provides a report-level fallback for the corresponding observation and assertion state. The same extraction procedure is applied to the sampled report Y^s, greedy report Y^g, and reference report Y^*.

Laterality and severity. Laterality and severity attributes are extracted from the sentence and anatomy spans associated with each matched finding. Deterministic lexicons are used to identify these attributes, with their normalized representation following Chest ImaGenome Wu et al. (2021). If no applicable laterality or severity attribute can be extracted, corresponding field is assigned $\emptyset$.

Multiple mentions and conflicting attributes. A report may mention the same normalized finding more than once. We therefore consolidate all mentions mapped to the same finding into a single structured state. Assertion states are resolved using the priority

$$\text{pos} > \text{unc} > \text{neg}. \tag{20}$$

For laterality, a finding is assigned bilateral status when both left and right mentions are present. When multiple severity descriptions conflict, the highest observed level among mild, moderate, and severe is retained. These rules ensure that every report yields at most one structured state for each finding category.

Special handling of positive findings. The *No Finding* category is excluded from positive-finding matching and from the five category-specific risk computations. We define a positive finding indicator as

$$\chi_k = \mathbf{1}[m_k = 1 \wedge z_k = \text{pos}]. \tag{21}$$

This indicator is used for positive-finding agreement and for the omission, unsupported-positive, laterality, and severity risk definitions in Section 3.2.

Uncertainty is treated as an assertion state rather than as a separate risk category. When a finding is mentioned in both generated and reference reports, a mismatch involving the uncertain state contributes to polarity inconsistency. Uncertainty-specific omission and unsupported-positive errors are not modeled separately.

Attribute eligibility. Laterality and severity are evaluated only when the corresponding attribute is specified by the reference report. If a generated report positively mentions such a finding but omits or mismatches the reference attribute, an attribute error is recorded. If the generated report omits the finding entirely, the discrepancy is counted as an omission rather than simultaneously as a laterality or severity error. This avoids double-counting a missing finding across multiple risk categories.

A.3 Derivation and Training Procedure of ACRA

Online risk statistics. For clarity, we first expand the mini-batch risk computation used in the online calibration step. For category c at iteration t, let

$$N_c^{(t)} = \sum_{i=1}^{B} \sum_{k=1}^{K} \omega_k \delta_{c,i}^k, \tag{22}$$

$$D_c^{(t)} = \sum_{i=1}^{B} \sum_{k=1}^{K} \omega_k a_{c,i}^k, \tag{23}$$

denote the total error mass and eligible mass, respectively, over the mini-batch of B greedy-reference pairs. The corresponding batch-level risk estimate is

$$b_c^{(t)} = \frac{N_c^{(t)}}{D_c^{(t)}}, \qquad D_c^{(t)} > 0. \tag{24}$$

Aggregating the numerator and denominator before forming their ratio is important because the number of eligible opportunities may differ substantially across reports and risk categories. Averaging report-level risks directly would also treat reports with no eligible opportunity as zero-risk observations, which would bias the estimate downward.

The online state is updated only when the current mini-batch contains at least one eligible opportunity:

$$\mu_c^{(t)} = \begin{cases} \rho \mu_c^{(t-1)} + (1-\rho) b_c^{(t)}, & D_c^{(t)} > 0, \\ \mu_c^{(t-1)}, & D_c^{(t)} = 0. \end{cases} \tag{25}$$

We initialize $\boldsymbol{\mu}^{(0)} = \mathbf{0}$. This update prevents a category with no evaluable instance in the current mini-batch from being interpreted as having zero clinical risk.

Derivation of the mirror-ascent update. Given the smoothed risk state $\boldsymbol{\mu}^{(t)}$, ACRA updates the allocation by solving

$$\boldsymbol{\alpha}^{(t)} = \arg\max_{\boldsymbol{\alpha} \in \Delta_{|\mathcal{C}|}} \left\{ \eta \langle \boldsymbol{\alpha}, \boldsymbol{\mu}^{(t)} \rangle - D_{\mathrm{KL}}\left(\boldsymbol{\alpha} \| \boldsymbol{\alpha}^{(t-1)}\right) - \eta\tau D_{\mathrm{KL}}\left(\boldsymbol{\alpha} \| \boldsymbol{\pi}\right) \right\}. \tag{26}$$

To derive the closed-form solution, we expand the KL terms and introduce a Lagrange multiplier ξ for the simplex constraint $\sum_c \alpha_c = 1$. Omitting terms independent of $\boldsymbol{\alpha}$, the Lagrangian is

$$\begin{aligned} \mathcal{J}(\boldsymbol{\alpha}, \xi) = \eta \sum_{c \in \mathcal{C}} \alpha_c \mu_c^{(t)} - \sum_{c \in \mathcal{C}} \alpha_c \log \frac{\alpha_c}{\alpha_c^{(t-1)}} \\ - \eta\tau \sum_{c \in \mathcal{C}} \alpha_c \log \frac{\alpha_c}{\pi_c} + \xi \left(\sum_{c \in \mathcal{C}} \alpha_c - 1 \right). \end{aligned} \tag{27}$$

Taking the derivative with respect to α_c and setting it to zero gives

$$\begin{aligned} 0 = \eta \mu_c^{(t)} - \left(\log \alpha_c - \log \alpha_c^{(t-1)} + 1 \right) \\ - \eta\tau \left(\log \alpha_c - \log \pi_c + 1 \right) + \xi. \end{aligned} \tag{28}$$

Algorithm 1 Sequence-level training with ACRA

Require: NLL-pretrained generator G_θ; training pairs (I, Y^*); prior $\boldsymbol{\pi}$; $\rho, \eta, \tau, \lambda, \gamma, \beta$

Initialize $\boldsymbol{\mu}^{(0)} \leftarrow \mathbf{0}$, $\boldsymbol{\alpha}^{(0)} \leftarrow \boldsymbol{\pi}$
for each mini-batch at iteration t **do**
  Generate sampled reports $Y^s \sim p_\theta(\cdot \mid I)$
  Generate greedy reports $Y^g = \mathrm{Greedy}(G_\theta, I)$
  Extract structured finding states for Y^s, Y^g, and Y^*
  for each risk category $c \in \mathcal{C}$ **do**
    Compute $N_c^{(t)}$ and $D_c^{(t)}$ from (Y^g, Y^*)
    if $D_c^{(t)} > 0$ **then**
      $b_c^{(t)} \leftarrow N_c^{(t)} / D_c^{(t)}$
      $\mu_c^{(t)} \leftarrow \rho \mu_c^{(t-1)} + (1-\rho) b_c^{(t)}$
    else
      $\mu_c^{(t)} \leftarrow \mu_c^{(t-1)}$
    end if
  end for
  Update $\boldsymbol{\alpha}^{(t)}$ using Eq. 31
  Compute $S^{(t)}(Y^s, Y^*)$ and $S^{(t)}(Y^g, Y^*)$ using the same $\mathrm{sg}[\boldsymbol{\alpha}^{(t)}]$
  $A^{(t)} \leftarrow S^{(t)}(Y^s, Y^*) - S^{(t)}(Y^g, Y^*)$
  Compute $\mathcal{L}_{\mathrm{seq}} = -\mathrm{sg}[A^{(t)}] \ell_\theta(Y^s \mid I)$
  Compute $\mathcal{L} = \beta \mathcal{L}_{\mathrm{NLL}} + (1-\beta) \mathcal{L}_{\mathrm{seq}}$
  Update θ using $\nabla_\theta \mathcal{L}$
end for

Rearranging,

$$(1+\eta\tau)\log\alpha_c = \log\alpha_c^{(t-1)} + \eta\mu_c^{(t)} + \eta\tau\log\pi_c + C, \tag{29}$$

where $C = \xi - (1+\eta\tau)$ is shared by all categories. Exponentiating both sides yields

$$\alpha_c \propto \left[\alpha_c^{(t-1)} \exp\left(\eta\mu_c^{(t)}\right) \pi_c^{\eta\tau}\right]^{\frac{1}{1+\eta\tau}}. \tag{30}$$

Normalizing over the simplex gives

$$\alpha_c^{(t)} = \frac{\left[\alpha_c^{(t-1)} \exp\left(\eta\mu_c^{(t)}\right) \pi_c^{\eta\tau}\right]^{\frac{1}{1+\eta\tau}}}{\sum_{j\in\mathcal{C}} \left[\alpha_j^{(t-1)} \exp\left(\eta\mu_j^{(t)}\right) \pi_j^{\eta\tau}\right]^{\frac{1}{1+\eta\tau}}}. \tag{31}$$

The update has two complementary effects. The exponential factor increases the mass assigned to categories with larger persistent risk, while the prior term pulls the allocation toward $\boldsymbol{\pi}$. The KL term relative to the previous allocation further limits abrupt changes across consecutive iterations. When $\tau = 0$, Eq. 31 reduces to the standard exponentiated-gradient update.

Training algorithm. Algorithm 1 summarizes the sequence-level training stage. The risk state is estimated from greedy decoding, whereas the sampled report provides the stochastic sequence used in the policy-gradient update. The updated allocation is then shared by the sampled and greedy reports so that their self-critical scores are evaluated under the same clinical objective.

Update ordering and gradient flow. The sequence-level update can be summarized as

$$\theta^{(t)} \to \{Y^s, Y^g\} \to \boldsymbol{\mu}^{(t)} \to \boldsymbol{\alpha}^{(t)} \to A^{(t)} \to \theta^{(t+1)}. \tag{32}$$

The greedy report is generated by the current model before the risk state and allocation are updated. The resulting $\boldsymbol{\alpha}^{(t)}$ is then held fixed for the corresponding generator update and is shared when evaluating Y^s and Y^g.

Neither the structured extractor nor the online risk-calibration procedure is differentiated. In particular, $\boldsymbol{\alpha}^{(t)}$ and $A^{(t)}$ are treated with stop-gradient in the generator update. The gradient with respect to θ therefore enters through the sampled sequence log-likelihood, while the clinical risk mechanism modulates its scalar self-critical advantage.

A.4 Additional Quantitative Results

A.4.1 Ablation Results on IU-Xray

We further evaluate the components of adaptive risk allocation on IU-Xray using the same ablation settings as in Table 5. *w/o Adaptive Allocation* fixes $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\pi}$; *w/o Prior Regularization* sets $\tau = 0$; and *w/o Risk Smoothing* replaces the EMA risk state with the current mini-batch estimate.

Table 6: Ablation of ACRA on IU-Xray. Higher is better for clinical consistency metrics and lower is better for reference-conditioned risks. Best values are shown in bold.

Variant	CheX-F1	RG-F1	F_{clin}	Avg.	Omi.	Unsup.	Pol.	Lat.	Sev.
w/o Adaptive Allocation	0.445	0.211	0.405	0.181	0.247	0.208	0.145	0.101	0.204
w/o Prior Regularization	0.449	0.214	0.411	0.170	**0.211**	**0.175**	0.141	0.104	0.219
w/o Risk Smoothing	0.451	0.216	0.414	0.168	0.224	0.188	0.137	0.099	0.192
ACRA	**0.454**	**0.218**	**0.419**	**0.162**	0.218	0.181	**0.132**	**0.093**	**0.186**

Table 6 shows a pattern consistent with the MIMIC-CXR ablation. Fixing the allocation increases R_{avg} from 0.162 to 0.181 and reduces all three clinical consistency metrics, providing additional evidence that adaptive allocation contributes beyond the use of structured clinical risks alone. Removing prior regularization increases R_{avg} to 0.170. Although omission and unsupported-positive risks decrease slightly, polarity, laterality, and particularly severity risk increase. This redistribution is consistent with the role of the prior in preventing the allocation from concentrating excessively on the dominant risk categories. Replacing the smoothed risk state with the current mini-batch estimate also degrades performance, increasing R_{avg} to 0.168 and reducing CheXbert-F1, RadGraph-F1, and F_{clin}. The full ACRA model therefore achieves the lowest average clinical risk together with the strongest overall clinical consistency on IU-Xray.

A.4.2 Additional Allocation Dynamics on IU-Xray

We further examine the training dynamics of ACRA on IU-Xray. Figure 4 shows the smoothed clinical risk state $\boldsymbol{\mu}^{(t)}$ and the corresponding adaptive allocation $\boldsymbol{\alpha}^{(t)}$ throughout sequence-level training.

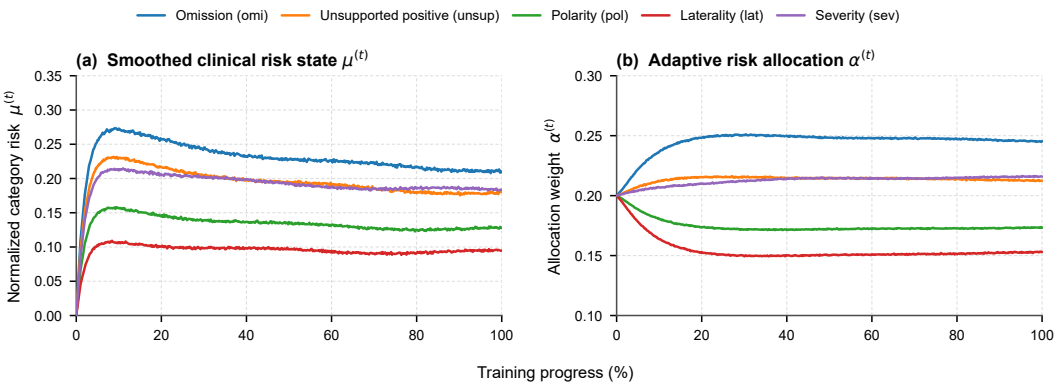


Figure 4: Training dynamics of the smoothed clinical risk state $\boldsymbol{\mu}^{(t)}$ (left) and adaptive risk allocation $\boldsymbol{\alpha}^{(t)}$ (right) on IU-Xray. All allocation weights are initialized from the uniform prior $\alpha_c^{(0)} = 0.2$.

As shown in Figure 4(a), the EMA state rapidly establishes the initial greedy-decoding risk profile and then decreases gradually as sequence-level training proceeds. The five risk categories remain clearly separated throughout most of training. Omission consistently exhibits the largest risk, while unsupported-positive and severity risks occupy an intermediate range. Polarity remains lower, and

laterality has the smallest residual risk. Although all categories decrease after the early stage, their relative magnitudes remain different, indicating that the clinical discrepancy profile is not well represented by a fixed uniform weighting. Figure 4(b) shows the corresponding evolution of the allocation. Starting from the uniform prior, ACRA rapidly reallocates mass toward the larger persistent risks. Omission receives the highest weight, while unsupported-positive and severity also receive increased emphasis. Conversely, polarity and laterality receive progressively smaller weights. The allocation stabilizes in the later stages of training rather than continuing to diverge, reflecting the combined effects of the proximal update and prior regularization. These dynamics are consistent with the behavior observed on MIMIC-CXR in Figure 3. Together with the IU-Xray ablation results in Table 6, they provide additional evidence that ACRA adapts its optimization focus to persistent category-specific risks while maintaining a distributed allocation across all five clinical error types.

A.5 Hyperparameter Sensitivity

We study the sensitivity of ACRA to three hyperparameters governing adaptive risk allocation: the EMA smoothing coefficient ρ, the mirror-ascent step size η, and the prior regularization strength τ. All experiments are conducted on MIMIC-CXR while varying one hyperparameter at a time and keeping the remaining settings fixed.

Table 7 shows that ACRA remains stable over a reasonably broad range of hyperparameter values. For the smoothing coefficient ρ, performance improves from $\rho = 0.70$ to 0.90 and slightly decreases at $\rho = 0.95$. This behavior is consistent with the role of EMA smoothing: insufficient smoothing leaves the risk state more sensitive to mini-batch variation, whereas excessive smoothing slows its response to changes in the generator. The method is also relatively insensitive to the mirror-ascent step size within a moderate range. Values of η between 0.05 and 0.20 produce similar clinical consistency and average risk, while a larger step size of 0.40 leads to a modest degradation. This suggests that the adaptive allocation does not depend on a narrowly tuned update rate. The prior regularization strength exhibits a clearer tradeoff. Weak regularization at $\tau = 0.10$ increases average risk, while overly strong regularization at $\tau = 2.00$ also degrades performance by constraining the allocation too closely to the prior. The intermediate setting $\tau = 0.50$ achieves the strongest overall results. These observations are consistent with the intended role of prior regularization in balancing responsiveness to persistent risks against excessive concentration of the allocation.

Table 7: Sensitivity of ACRA to the risk smoothing coefficient ρ, mirror-ascent step size η, and prior regularization strength τ on MIMIC-CXR. Higher is better for clinical consistency metrics and lower is better for R_{avg}. The default setting is shown in bold.

Parameter	Value	CheX-F1	RG-F1	F_{clin}	R_{avg}
ρ	0.70	0.397	0.209	0.377	0.181
	0.80	0.398	0.210	0.379	0.179
	0.90	**0.399**	**0.211**	**0.381**	**0.177**
	0.95	0.398	0.210	0.380	0.179
η	0.05	0.398	0.210	0.379	0.179
	0.10	**0.399**	**0.211**	**0.381**	**0.177**
	0.20	0.398	0.210	0.380	0.178
	0.40	0.397	0.209	0.377	0.181
τ	0.10	0.396	0.208	0.375	0.183
	0.50	**0.399**	**0.211**	**0.381**	**0.177**
	1.00	0.398	0.210	0.379	0.180
	2.00	0.395	0.208	0.374	0.184