CORRECT BEFORE IT DRIFTS: COUNTERFACTUAL-GUIDED INTERVENTION TO MITIGATE HALLUCINATIONS IN RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology-report hallucinations are usually judged after generation, although an autoregressive generator commits to factual statements sequentially. Once an image-inconsistent statement enters the partial report, it becomes context for what follows. We call this failure mode *partial-report drift* and ask a specific online question: when an image-grounded repair is available, should it be applied now? **Correct Before It Drifts** first verifies newly generated claims and proposes minimal supported edits, then uses matched corrected/unchanged complete reports as offline preference supervision for a lightweight short-lookahead scorer. At test time, the scorer compares only short continuations and applies a repair when its margin clears a validation-calibrated threshold. Across MIMIC-CXR and IU-Xray, controlled generators, inference-time baselines, and matched generator-adaptation settings, the method improves fine-grained clinical factuality with modest overhead. Short lookahead predicts complete-report preference better than local verification, and early repairs have substantially larger suffix-level effects than late repairs, supporting hallucination mitigation as online state correction rather than only post-hoc revision.

1 INTRODUCTION

Radiology report generation is increasingly built on medical vision–language models that combine domain-specific pretraining, instruction tuning, and visual grounding (Chen et al., 2024; Zambrano Chaves et al., 2025; Bannur et al., 2024). Despite substantial progress, clinically meaningful hallucinations remain common: a report may introduce an unsupported finding, invert a negation, attach the wrong anatomical modifier, or add an unsupported descriptive attribute. Most evaluation and correction pipelines operate on the completed report. Autoregressive generation, however, creates a different failure surface: every sentence becomes conditioning context for the next one.

This sequential dependence matters because an early error can change what follows. Multimodal hallucinations can snowball through subsequent generation (Zhong et al., 2024), and sentence-level early-intervention work shows that hallucinations often emerge before the final output is complete (Peng et al., 2025). We refer to the corresponding radiology setting as *partial-report drift*: a factual error in the retained prefix perturbs the distribution of later report content. Figure 1 illustrates the core distinction. Correcting one sentence can do more than fix that sentence; it can also prevent later text from being generated from a flawed state.

Existing methods address adjacent but different points in this pipeline. Generic inference-time approaches modify token probabilities, visual reliance, or attention during decoding (Leng et al., 2024; Huang et al., 2024; Li et al., 2025); CCD injects radiology-specific clinical signals into contrastive decoding (Zhang et al., 2026). Other methods update the report generator itself through grounding-aware training, self-correction, or reinforcement learning (Park et al., 2025; Messina et al., 2026; Liu et al., 2026; Chen et al., 2026; Zhang et al., 2025). Post-hoc systems verify or rewrite unsupported content after generation (Mahmood et al., 2025; Heiman et al., 2025; Zhou et al., 2026; Mahmood & Syeda-Mahmood, 2026). None directly isolates the decision that arises at a live partial-report state: *given one concrete image-supported edit to the current sentence, is it better to apply that edit before the same generator continues?*

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.

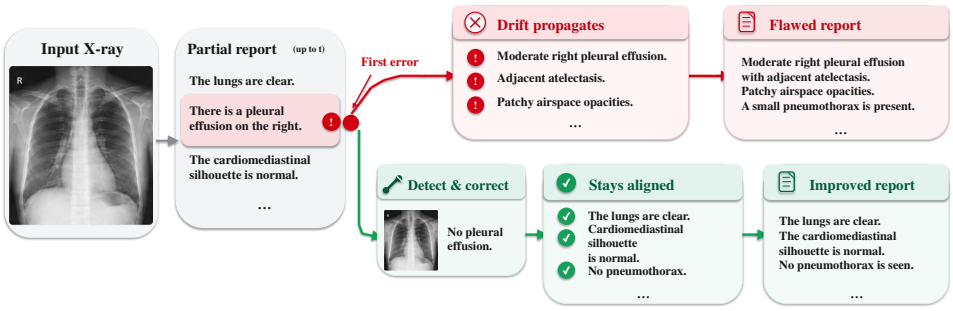


Figure 1: **Motivation.** A local image-inconsistent statement becomes part of the autoregressive context. Correcting it before subsequent sentences are generated can prevent the report from drifting further from the image.

The distinction is important because *repair validity is not repair utility*. A local edit may be image-supported yet produce no meaningful downstream benefit, or may perturb an otherwise acceptable continuation. Conversely, keeping an erroneous state can allow later sentences to inherit the error. We therefore separate the problem into two questions: **what can be safely repaired from the image?** and **whether applying that repair now improves the downstream report?**

We propose **Correct Before It Drifts**. Grounded Drift Detection (GDD) verifies newly introduced claims and proposes a minimal supported repair. Counterfactual Preference Learning (CPL) creates matched corrected and unchanged branches from the same prefix, continues both with the same generator, and uses complete-report outcomes only as offline supervision. A compact preference model then learns to predict that complete-outcome ordering from a short lookahead. Online Repair Selection (ORS) deploys the learned preference with a calibrated threshold and reuses the selected lookahead, keeping the online cost bounded.

Our contributions are threefold. First, we formulate partial-report drift as a testable online state-correction problem rather than a synonym for generic hallucination. Second, we introduce a matched counterfactual learning scheme that transfers complete-report supervision to a short-lookahead selector without updating the report generator. Third, we evaluate both final report quality and the proposed mechanism: grounding reliability, short-lookahead preference prediction, suffix-level drift, cross-dataset and cross-generator transfer, complementarity with generator adaptation, selective execution, and runtime.

2 Related Work

Radiology generation and factuality. CheXagent, LLaVA-Rad, and MAIRA-2 improve radiology generation through domain-specific training and grounding (Chen et al., 2024; Zambrano Chaves et al., 2025; Bannur et al., 2024). DART, CURE, SAT-RRG, OraPO, and EditGRPO further improve factuality by re-alignment, grounding-aware curricula, localized optimization, or clinically motivated rewards/edits (Park et al., 2025; Messina et al., 2026; Liu et al., 2026; Chen et al., 2026; Zhang et al., 2025). These methods update the generator; our main setting intervenes on residual errors while leaving it fixed.

Verification, decoding, and correction. VCD, OPERA, VISTA, and CCD mitigate hallucination by altering decoding or visual information flow (Leng et al., 2024; Huang et al., 2024; Li et al., 2025; Zhang et al., 2026). Radiology-specific verification/correction includes phrase-grounded fact checking, FactCheXcker, CoEV, and phrase-grounded APO (Mahmood et al., 2025; Heiman et al., 2025; Zhou et al., 2026; Mahmood & Syeda-Mahmood, 2026); Radiology PRMs score sentence-level factuality for filtering or selection (Thomas et al., 2025). Our target is different: whether changing the current state improves what the same generator produces next.

Early intervention and process supervision. Hallucination snowballing and SENTINEL motivate attention to where an error first enters a sequence (Zhong et al., 2024; Peng et al., 2025); SENTINEL, however, learns early intervention into model parameters rather than making a live repair

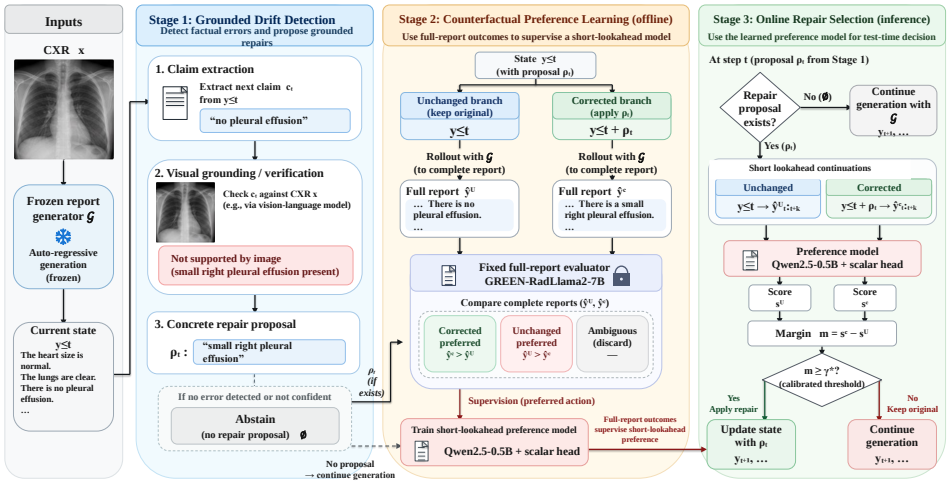


Figure 2: **Framework.** Stage 1 proposes a concrete image-grounded repair. Stage 2 constructs matched corrected and unchanged complete reports offline and converts their full-report outcomes into pairwise preference supervision. Stage 3 uses the same learned scorer on short H-sentence continuations and applies the repair only when the corrected branch clears a calibrated margin.

decision. Process-supervision work shows that terminal outcomes can supervise intermediate decisions (Jia et al., 2025; Setlur et al., 2025); we use matched complete-report outcomes to supervise a short-lookahead intervention model.

3 METHOD

Let x be the radiograph, c the available clinical context, G a frozen report generator, and $y_{\le t} = [u_1, \ldots, u_t]$ the report after the t-th complete sentence. Figure 2 summarizes the three stages. The design deliberately separates three quantities that are easy to conflate. GDD asks whether the image supports a concrete local edit; CPL asks which of two matched report states leads to the better complete outcome; ORS asks whether the evidence available from a short continuation is strong enough to commit that edit online. This separation is important: the visual module is never asked to predict long-range language dynamics, while the preference model is never allowed to invent a correction that has not already been grounded in the image.

3.1 GROUNDED DRIFT DETECTION

Structured claim extraction. We parse each new sentence with frozen RadGraph-XL (Delbrouck et al., 2024). The primary clinical assertion is represented as

$$c_j = (f_j, p_j, \ell_j), \tag{1}$$

where f is a finding, $p \in \{\text{present}, \text{absent}, \text{uncertain}\}$ is its assertion state, and ℓ is an anatomical location when available. A running report graph caches previously processed assertions so that only new or modified content is re-evaluated. Detachable descriptive modifiers are retained as aligned spans; the repair operator may delete an unsupported modifier but never invent an unverified severity, measurement, or temporal comparison.

Image-grounded verification. We use frozen BioViL-T image/text encoders (Bannur et al., 2023). The image is encoded once,

$$(\mathbf{z}_x, \mathbf{P}_x) = E_I(x), \tag{2}$$

and cached. Each claim is converted to a deterministic canonical phrase $q(c)$ and embedded as $\mathbf{z}_c = E_T(q(c))$. A lightweight trainable verifier V_ϕ combines the global image vector, claim-conditioned spatial features, and the text embedding to produce an image-support score $s_\phi(x, c)$.

It is trained on real and synthetically perturbed finding–location pairs following phrase-grounded fact checking (Mahmood et al., 2025).

Two thresholds τ_- and τ_+ are selected on patient-disjoint validation data. Claims with $s_\phi \geq \tau_+$ are supported, those with $s_\phi \leq \tau_-$ unsupported, and intermediate cases are unresolved. Thus ambiguous visual evidence causes abstention rather than forced correction.

Minimal repair proposal. For a high-confidence discrepancy, GDD returns at most one structured repair $\rho_t : u_t \to \widetilde{u}_t$. Polarity is flipped only if the opposite assertion is independently supported. Unsupported positive findings or detachable unsupported modifiers can be removed. An anatomical modifier can be removed when the location-free assertion remains supported, and can be replaced only when a unique alternative location is supported. The repaired sentence is parsed again; if a non-target clinical assertion changes, the edit is rejected. Pure omissions that are never expressed in the report are outside the core online repair scope.

3.2 Counterfactual Preference Learning

A valid local repair need not be useful at the current state. We therefore create matched alternatives that differ only in whether the current repair is applied:

$$y^U_{\leq t} = [u_1, \ldots, u_t], \tag{3}$$

$$y^C_{\leq t} = [u_1, \ldots, \widetilde{u}_t]. \tag{4}$$

The same generator, context, and decoding configuration roll both states to EOS,

$$Y^a_t = G_{\to \mathrm{EOS}}(x, c, y^a_{\leq t}), \qquad a \in \{C, U\}. \tag{5}$$

No later repair is executed while constructing this pair, so the active repair is the only intervention-level difference. Counterfactual here is operational: the branches are matched model-generated alternatives, not a claim of causal identification.

Conservative complete-report supervision. We use fixed GREEN-RadLlama2-7B (Ostmeier et al., 2024) only during offline pair construction. For candidate report Y and radiologist reference r, let $M(Y, r)$ be the number of matched findings and $E(Y, r)$ the number of clinically significant GREEN errors. We prefer the corrected branch if

$$E(Y^C_t, r) \leq E(Y^U_t, r), \quad M(Y^C_t, r) \geq M(Y^U_t, r), \tag{6}$$

with at least one strict inequality; the reverse relation defines $U \succ C$. Error–coverage trade-offs and ties are discarded. This dominance rule avoids hand-designing a scalar exchange rate between completeness and error count. GREEN and the reference are never available online.

Preference corpus and corrected-prefix states. Counterfactual data are built entirely offline, before the online selector exists. We start from 16,700 patient-disjoint MIMIC-CXR training studies and generate one report with each of CheXagent, LLaVA-Rad, and MAIRA-2, producing 50,100 initial report instances. GDD identifies 24,386 parser-level discrepancy states; 13,642 survive visual verification and repair-validity checks and are expanded into matched branches. Conservative dominance retains 8,714 training pairs: 5,394 (61.9%) prefer correction and 3,320 (38.1%) prefer keeping the original state. The sizeable unchanged-preferred fraction is direct evidence that a grounded repair is not automatically a useful intervention. After a pair is labeled, its generated branches are scanned for later proposals, so later training states can contain earlier repairs in their prefix without any cold-start dependence on a pretrained selector. Details are in Appendix B.

Short-lookahead preference model. Full rollouts provide supervision but are too expensive for every online decision. For horizon H complete sentences after the decision sentence,

$$S^a_{t,H} = y^a_{\leq t} \oplus [u^a_{t+1}, \ldots, u^a_{t+H}], \tag{7}$$

using the available suffix if EOS arrives earlier. We initialize a shared scorer f_θ from Qwen2.5-0.5B (Yang et al., 2024) and add a scalar head. The scorer receives clinical context, the common

prefix, the candidate decision sentence, and its branch-specific lookahead; it receives neither generator identity nor an explicit corrected/unchanged token. Scores

$$r^a_{t,H} = f_\theta(c, S^a_{t,H}) \tag{8}$$

induce

$$P_\theta(C \succ U) = \sigma(r^C_{t,H} - r^U_{t,H}). \tag{9}$$

We optimize the Bradley–Terry pairwise loss (Bradley & Terry, 1952; Sun et al., 2025). The main system uses $H = 2$; Sec. 4.5 and Appendix B analyze $H \in \{0, 1, 2, 3\}$.

3.3 Online Repair Selection

At test time, if GDD abstains, generation continues unchanged. Otherwise, the frozen generator produces up to H additional sentences from both branches and the scorer computes the preference margin

$$m_t = f_\theta(c, S^C_{t,H}) - f_\theta(c, S^U_{t,H}). \tag{10}$$

A positive margin alone is not enough to edit the report. On held-out preference pairs we select

$$\gamma^* = \arg\max_\gamma C_{\text{app}}(\gamma) \quad \text{s.t.} \quad P_{\text{app}}(\gamma) \geq \pi_{\min}, \tag{11}$$

where $P_{\text{app}}(\gamma) = P(z = C \mid m \geq \gamma)$ is selected-repair precision and $C_{\text{app}}(\gamma) = P(m \geq \gamma)$ is coverage. We use $\pi_{\min} = 0.90$. The repair is applied iff $m_t \geq \gamma^*$. This calibration makes the role of ORS explicit: it is not designed to maximize the number of edits, but to expose a controllable precision–coverage operating point for edits that alter a live report. The selected lookahead is buffered and reused rather than regenerated. Buffered sentences remain subject to the same sentence-level detector; if one later triggers a new proposal, the stale remainder was generated from a prefix that may no longer be retained, so it is discarded and a fresh corrected/unchanged comparison begins.

4 Experiments

4.1 Experimental Setup

Evaluation questions. The experiments are organized around six questions. **Q1:** Does selective online intervention improve factual report quality under a controlled matched-generator comparison? **Q2:** Does the learned corrected-versus-unchanged preference transfer across datasets and unseen report generators? **Q3:** Does online intervention remain useful after the generator itself has been improved through factuality-oriented training? **Q4:** Do the intermediate measurements support the proposed mechanism—reliable grounded proposals, useful short lookahead, and a larger downstream effect for earlier repairs? **Q5:** Which stages are necessary for proposal safety, downstream awareness, and conservative execution? **Q6:** What distinguishes an intervention that is applied from a grounded proposal that is deliberately kept unchanged? This structure makes each experiment answer a specific claim rather than treating the evaluation as a flat collection of metrics.

Datasets, generators, and controlled protocol. MIMIC-CXR (Johnson et al., 2019) is the primary benchmark and source of counterfactual preference data; IU-Xray (Demner-Fushman et al., 2016) is reserved for zero-shot cross-dataset transfer. The controlled MIMIC-CXR AP/PA Findings cohort contains 162,969/1,286/2,461 train/validation/test studies, and IU-Xray uses a fixed 2,769/395/791 split. Preference learning, verifier training, and threshold calibration are patient-disjoint from the final MIMIC test set. The controlled comparison uses CheXagent (Chen et al., 2024), LLaVA-Rad (Zambrano Chaves et al., 2025), and MAIRA-2 (Bannur et al., 2024). Within each generator block, every method receives the same radiographs, clinical context, generator checkpoint, and base decoding configuration; disabling a mitigation-specific adapter must recover the corresponding base output.

Baselines, metrics, and supervision/evaluation separation. We compare with VCD (Leng et al., 2024), OPERA (Huang et al., 2024), VISTA (Li et al., 2025), and CCD (Zhang et al., 2026), spanning visual contrast, context-overtrust mitigation, visual-information steering, and radiology-specific clinical decoding. Headline metrics are CREPE (Cho et al., 2025) (predicted clinical error

Table 1: **MIMIC-CXR hallucination mitigation.** Methods within each block share the same generator; latency is relative to Base. Best/second-best values are bold/underlined.

Generator	Method	CREPE↓	RaTE↑	RadG↑	CXB14↑	BERT↑	Lat.× ↓
CheXagent	Base	1.84	52.6	26.6	55.2	84.9	1.00
	VCD	1.79	53.1	26.0	56.0	84.6	1.86
	OPERA	1.77	53.4	27.1	56.5	**85.2**	1.63
	VISTA	1.72	54.0	27.3	57.0	85.1	1.31
	CCD	1.68	54.8	27.6	**58.1**	84.8	1.74
	Ours	**1.57**	**56.1**	**28.4**	57.8	85.1	**1.24**
LLaVA-Rad	Base	1.96	50.8	24.9	57.3	83.2	1.00
	VCD	1.91	51.4	24.4	57.0	82.9	1.84
	OPERA	1.88	51.7	25.3	57.8	**83.7**	1.61
	VISTA	1.84	52.3	25.5	58.5	83.6	1.30
	CCD	1.80	53.0	25.9	**59.4**	83.3	1.73
	Ours	**1.69**	**54.6**	**26.8**	59.0	83.6	**1.28**
MAIRA-2	Base	1.46	58.4	34.6	58.1	86.1	1.00
	VCD	1.43	58.8	34.1	58.5	85.9	1.81
	OPERA	1.41	59.0	34.9	58.0	**86.5**	1.59
	VISTA	1.37	59.6	35.2	59.4	86.4	1.28
	CCD	1.34	60.3	35.5	**60.4**	86.0	1.72
	Ours	**1.26**	**61.4**	**36.2**	60.0	86.4	**1.22**

count; lower is better), RaTEScore (Zhao et al., 2024), RadGraph F1 (Jain et al., 2021), CheXbert-14 Micro-F1 (Smit et al., 2020), and BERTScore (Zhang et al., 2020); all except CREPE are multiplied by 100. GREEN is used only to construct offline pairwise preferences on training data and is never available at inference or used as a headline test metric. Principal comparisons are paired by study with 1,000 paired-bootstrap replicates; preference training is repeated with three seeds while reusing cached continuations.

4.2 Controlled Hallucination Mitigation

Explicit repair is most consistently reflected in fine-grained clinical metrics. Relative to the unmodified generators, Ours reduces CREPE from 1.84 to 1.57, 1.96 to 1.69, and 1.46 to 1.26 for CheXagent, LLaVA-Rad, and MAIRA-2, respectively, while improving RaTEScore by +3.5/+3.8/+3.0 and RadGraph by +1.8/+1.9/+1.6 points. It also achieves the lowest CREPE and highest RadGraph in every generator block. These gains span report-level error counts, entity-aware semantics, and structured relations, matching the intended role of explicit grounded repair rather than a single disease-label proxy.

Metric trade-offs are consistent with what each method optimizes. CCD remains particularly strong on CXB14 because its radiology expert directly injects disease-level signals during decoding, while OPERA/VISTA can be stronger on BERTScore because they reshape generation without explicitly replacing a clinical span. Ours need not maximize either proxy: correcting an image-inconsistent phrase can improve clinical content while slightly changing reference wording. The important pattern is that CREPE, RaTEScore, and RadGraph improve together while BERTScore is essentially preserved.

Clinical gains do not require continuous branching. Our additional branch is created only after GDD proposes an actionable repair, and the selected lookahead is reused as the live continuation. End-to-end latency is therefore 1.22–1.28×, compared with 1.72–1.86× for the continuously modified VCD/CCD configurations in this controlled setup. The result shows that downstream-aware selection can be invoked sparsely rather than maintaining a second decoding stream throughout the report.

4.3 Cross-Dataset and Cross-Generator Generalization

Cross-dataset transfer tests dependence on MIMIC-specific report conventions. The preference model, lookahead horizon, grounding thresholds, and intervention threshold are frozen before evaluation on IU-Xray. The gains are smaller than on MIMIC-CXR, as expected under shifts in report style and finding distribution, but CREPE, RaTEScore, and RadGraph improve for both evaluated generators. CCD remains competitive and is stronger on CXB14, so the transfer result is not produced by uniformly favoring our method across metrics.

Table 2: **Generalization.** (a) MIMIC-CXR → IU-Xray with no update/recalibration. (b) Leave-one-generator-out preference training.

Setting / Generator	Method	CREPE↓	RaTE↑	RadG↑	CXB14↑	BERT↑
(a) MIMIC-CXR → IU-Xray						
LLaVA-Rad	Base	2.14	47.5	22.8	51.8	81.7
	VISTA	2.01	48.9	23.5	52.8	**82.4**
	CCD	1.93	49.8	24.0	**53.8**	82.0
	Ours	**1.82**	**51.1**	**24.8**	53.5	82.3
MAIRA-2	Base	1.82	50.4	27.1	52.5	82.8
	VISTA	1.73	51.7	27.7	53.6	**83.3**
	CCD	1.68	52.6	28.1	**54.6**	82.9
	Ours	**1.57**	**54.0**	**28.9**	54.2	83.2
(b) Leave-one-generator-out on MIMIC-CXR						
CheXagent	Base	1.84	52.6	26.6	55.2	84.9
	Ours (train L+M)	**1.64**	**55.0**	**27.9**	**57.1**	**85.0**
LLaVA-Rad	Base	1.96	50.8	24.9	57.3	83.2
	Ours (train C+M)	**1.76**	**53.4**	**26.1**	**58.5**	**83.4**
MAIRA-2	Base	1.46	58.4	34.6	58.1	86.1
	Ours (train C+L)	**1.33**	**60.0**	**35.5**	**59.3**	**86.2**

Leave-one-generator-out evaluation tests a different failure mode. Here the target generator contributes no preference-training pair. Performance remains above the corresponding base generator, but the gains are weaker than with pooled training. This degradation is informative: the selector benefits from generator-specific continuation statistics, yet a substantial part of the corrected-versus-unchanged preference transfers across generators. The learned decision rule is therefore neither purely generator-specific nor perfectly generator-invariant.

4.4 Beyond Parameter-Frozen Generators

The controlled setting isolates inference-time intervention, but it does not show whether online repair remains useful after the generator itself has already been optimized for factuality. We therefore perform matched-generator comparisons with two generator-adaptation paradigms; Ours is attached to the same base or adapted generator rather than compared across unrelated model families.

Table 3: **Matched-generator comparison with generator adaptation.** "Upd." denotes generator parameter updates; "Online" denotes an added test-time intervention module.

Family	Method	Upd.	Online	CREPE↓	RaTE↑	RadG↑	CXB14↑	BERT↑
MedGemma-4B	Base	N	N	1.71	54.2	29.6	54.8	84.1
	+ Ours	N	Y	1.49	56.6	31.0	56.2	84.3
	CURE (Messina et al., 2026)	Y	N	1.38	57.8	32.1	57.4	**84.8**
	CURE + Ours	Y	Y	**1.29**	**58.7**	**32.8**	**58.1**	84.7
Qwen2.5-VL-3B	SFT	Y	N	1.92	51.8	27.2	52.8	83.4
	SFT + Ours	Y	Y	1.61	54.8	29.5	55.4	83.7
	EditGRPO (Zhang et al., 2025)	Y	N	1.54	55.5	30.0	56.0	**83.9**
	EditGRPO + Ours	Y	Y	**1.43**	**56.4**	**30.8**	**56.8**	**83.9**

Strong generator training is not replaced by inference-time intervention. CURE is stronger than attaching Ours to the unadapted MedGemma base on the main clinical metrics. This is expected: CURE explicitly trains anatomy-grounded report generation, whereas our method can only act on residual errors after they appear. The comparison prevents interpreting parameter-frozen intervention as a substitute for strong generator training.

Residual intervention headroom remains after adaptation. Adding Ours on top of CURE improves CREPE/RaTEScore/RadGraph from 1.38/57.8/32.1 to 1.29/58.7/32.8; similarly, EditGRPO+Ours improves the corresponding Qwen-family metrics from 1.54/55.5/30.0 to 1.43/56.4/30.8. The gain is smaller after adaptation, consistent with fewer residual errors, but it does not disappear. Training-time factuality optimization and test-time state correction therefore address partially distinct failure modes.

4.5 Process-Level Validation

Final-report tables establish usefulness but not the mechanism claimed in Sec. 3. We therefore test three links separately: whether grounded verification is sufficiently reliable to define a conservative action set, whether short downstream context predicts the complete-outcome preference better than

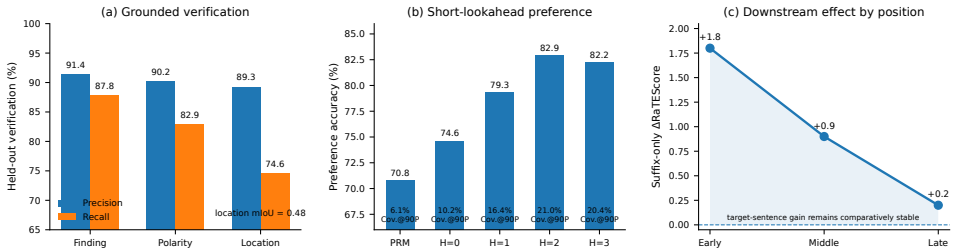


Figure 3: **Process validation.** (a) Held-out grounded verification. (b) Full-outcome preference prediction across lookahead horizons, with Cov.@90P. (c) Suffix-only RaTEScore gain after excluding the repaired sentence.

Table 4: **Cumulative ablation (MAIRA-2).** Apply-P: repair precision; Intv.: execution rate.

Variant	CREPE↓	RaTE↑	RadG↑	CXB14↑	BERT↑	Apply-P↑	Intv.%
Base generator	1.46	58.4	34.6	58.1	86.1	–	0
Grounded repair only	1.43	58.6	34.9	58.7	86.1	73.5	46.2
Local selection ($H = 0$)	1.39	59.1	35.2	59.0	**86.6**	84.6	18.9
Lookahead ($H = 2, m > 0$)	1.31	60.5	35.9	59.7	86.3	82.1	30.8
Full selective lookahead	**1.26**	**61.4**	**36.2**	**60.0**	86.4	**91.2**	21.5

local verification, and whether early repair changes the subsequent suffix rather than only the edited sentence.

Grounding is reliable enough for selective repair, with localization as the limiting case. Figure 3(a) evaluates the visual verifier independently of the final report generator. Finding and polarity verification reach 91.4% and 90.2% precision, while location reaches 89.3% precision but lower recall (74.6%) and only 0.48 localization mIoU. This gap motivates the asymmetric repair rules: an unsupported location can be removed conservatively, whereas a replacement location is introduced only when a unique alternative is independently supported. GDD is therefore a high-precision proposal mechanism, not a complete radiology interpreter.

The useful future signal is short-range rather than arbitrarily long. A sentence-level Radiology PRM (Thomas et al., 2025) reaches 70.8% full-outcome preference accuracy. The matched $H = 0$ scorer reaches 74.6%, and one/two future sentences increase accuracy to 79.3/82.9%; at the same 90% selected-repair precision, coverage rises from 10.2% at $H = 0$ to 21.0% at $H = 2$. $H = 3$ slightly drops to 82.2% while increasing latency to 1.30×. Thus the benefit comes from observing the generator's immediate response to the changed state, and it saturates rather than growing automatically with rollout length.

Early repair changes the suffix, not only the repaired sentence. After excluding the intervention sentence from both branches, the suffix-only RaTEScore gain decreases from +1.8 for early interventions to +0.9 for middle and +0.2 for late interventions, while the target-sentence benefit remains comparatively stable across position. This separation is what the partial-report drift hypothesis predicts: an early state has more remaining generation through which its effect can propagate, whereas a late repair increasingly behaves like an isolated local edit. Without this suffix-only component, the method would be better characterized as online correction rather than drift mitigation.

4.6 Ablation Study

Grounded repair is useful, but direct execution is too aggressive. Applying every high-confidence proposal already improves the base generator, confirming that Stage 1 identifies useful corrections. Its 73.5 Apply-P and 46.2% intervention rate, however, show that visual support alone does not determine whether the live report should be modified. This isolates the need for repair selection rather than conflating proposal validity with downstream utility.

Local selection improves reliability but leaves clinical benefit on the table. At $H = 0$, the same pairwise formulation substantially raises Apply-P while using no future sentence. The remaining gap

to $H = 2$ on CREPE, RaTEScore, and RadGraph shows that the current sentence alone does not contain all information needed to estimate downstream utility.

Lookahead supplies most of the gain; calibration controls how aggressively it is used. The unthresholded $H = 2$ selector executes more marginal repairs and reaches only 82.1 Apply-P. The full selector preserves the lookahead benefit while raising Apply-P to 91.2 and reducing intervention rate to 21.5%. The calibrated threshold is therefore a reliability mechanism rather than the source of the downstream signal itself.

4.7 Qualitative Intervention Analysis

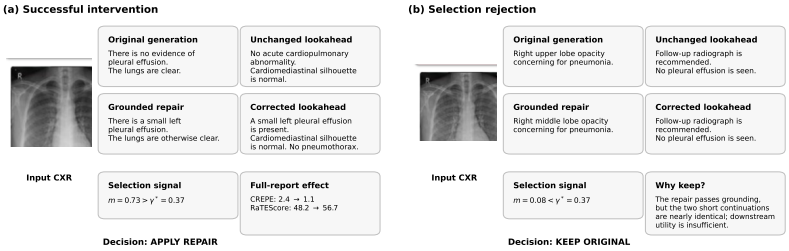


Figure 4: **APPLY versus KEEP.** A high-margin repair changes the live state; a grounded but low-utility proposal is retained unchanged.

Figure 4 makes the policy concrete: the APPLY case clears γ^* and improves CREPE 2.4→1.1 and RaTEScore 48.2→56.7, whereas the grounded KEEP proposal has nearly identical $H = 2$ continuations and $m = 0.08 < \gamma^*$. The policy therefore distinguishes *can be repaired* from *is useful enough to repair now*.

5 Limitations

The action space is intentionally conservative: pure omissions, unsupported severity, measurements, device-depth statements, and ambiguous fine-grained locations are not repaired; errors missed by RadGraph-XL or the verifier are unreachable. GREEN-derived complete-outcome preferences are an imperfect supervisor despite conservative filtering and independent test metrics. Actionable states also incur extra decoding and a modified prefix; transfer tests robustness empirically but cannot guarantee it for arbitrary future generators.

6 Conclusion

Correct Before It Drifts separates *what can be safely repaired* from *whether repairing it now improves what follows*. Complete corrected/unchanged futures supervise that decision offline, while short lookahead makes it online. Gains across generators, datasets, and adapted models, together with the suffix-only early-intervention effect, support selective state correction as a complement to generator training and post-hoc revision.

Reproducibility Statement

We specify the exact report generators, data cohorts, verifier construction, counterfactual-pair collection, preference objective, selection rule, metric implementations, statistical protocol, and runtime measurements in Appendix A–C. The online algorithm is given in Appendix B; all test-time comparisons use the same base generator inputs and decoding configuration within each generator block.

AI Use Statement

In this work, generative AI tools, including ChatGPT, were used to assist with research ideation and refinement, feedback on methodological and experimental design, literature discovery and analysis, and drafting, restructuring, and language editing of portions of the manuscript. All methodological choices, experiments, implementations, analyses, references, and claims were independently reviewed and validated by the authors. References suggested through AI-assisted literature discovery were checked against original sources, and all AI-assisted manuscript content was manually reviewed and revised. The authors take full responsibility for the final content of the work.

References

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15016–15027, 2023.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, et al. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Ralph A. Bradley and Milton E. Terry. The rank analysis of incomplete block designs. i. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Zhihong Chen, Maya Varma, Justin Xu, et al. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

Zhuoxiao Chen, Hongyang Yu, Ying Xu, Yadan Luo, Long Duong, and Yuan-Fang Li. OraPO: Oracle-educated reinforcement learning for data-efficient and factual radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28275–28287, 2026.

Gihun Cho, Seunghyun Jang, Hanbin Ko, Inhyeok Baek, and Chang Min Park. CREPE: Rapid chest x-ray report evaluation by predicting multi-category error counts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21738–21755. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1102.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12902–12915. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.765.

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. doi: 10.1093/jamia/ocv080.

Alice Heiman, Xiaoman Zhang, Emma Chen, Sung Eun Kim, and Pranav Rajpurkar. FactCheXcker: Mitigating measurement hallucinations in chest x-ray report generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30787–30796, 2025. doi: 10.1109/CVPR52734.2025.02867.

Qidong Huang, Xiaoyi Dong, Pan Zhang, et al. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, et al. RadGraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Do we need to verify step by step? rethinking process supervision from a theoretical perspective. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27373–27398. PMLR, 2025.

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*, 6:317, 2019. doi: 10.1038/s41597-019-0322-0.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819. PMLR, 2025.

Yunyi Liu, Yingshu Li, Tong Chen, Lingqiao Liu, Lei Wang, and Luping Zhou. SAT-RRG: Llm-guided self-adaptive training for radiology report generation with token-level push-pull optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35363–35372, 2026.

Razi Mahmood, Diego Machado-Reyes, Joy Wu, Parisa Kaviani, Ken C. L. Wong, Niharika D'Souza, Mannudeep Kalra, Ge Wang, Pingkun Yan, and Tanveer Syeda-Mahmood. Phrase-grounded fact-checking for automatically generated chest x-ray reports. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15966 of *Lecture Notes in Computer Science*, pp. 441–452. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-04981-0_42.

Raziuddin Mahmood and Tanveer Syeda-Mahmood. Phrase-grounded APO for improving chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28254–28263, 2026.

Pablo Messina, Andrés Villa, Juan Leon Alcazar, Karen Sanchez, Carlos Hinojosa, Denis Parra, Alvaro Soto, and Bernard Ghanem. CURE: Curriculum-guided multi-task training for reliable anatomy grounded report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36279–36289, 2026.

Sophie Ostmeier, Justin Xu, Zhihong Chen, et al. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.21.

Sang-Jun Park, Keun-Soo Heo, Dong-Hee Shin, Young-Han Son, Ji-Hye Oh, and Tae-Eui Kam. DART: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15580–15589, 2025.

Shangpin Peng, Senqiao Yang, Li Jiang, and Zhuotao Tian. Mitigating object hallucinations via sentence-level early intervention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 635–646, 2025.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. In *International Conference on Learning Representations*, 2025.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew 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.

Hao Sun, Yunyi Shen, and Jean-Francois Ton. Rethinking reward modeling in preference-based large language model alignment. In *International Conference on Learning Representations*, 2025.

Alois Thomas, Maya Varma, Jean-Benoit Delbrouck, and Curtis P. Langlotz. Process reward models for sentence-level verification of lvlm radiology reports. *arXiv preprint arXiv:2510.23217*, 2025.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Juan Manuel Zambrano Chaves et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16:3108, 2025. doi: 10.1038/s41467-025-58344-x.

Kai Zhang, Christopher Malon, Lichao Sun, and Martin Renqiang Min. EditGRPO: Reinforcement learning with post-rollout edits for clinically accurate chest x-ray report generation. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 304–316. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-short.26.

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

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. CCD: Mitigating hallucinations in radiology mllms via clinical contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35174–35200. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1755.

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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.836.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 11991–12011. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.648.

Nan Zhou, Ke Zou, Meng Liu, Linchao He, Jiaqi Zhu, Yi Zhang, Hu Chen, and Huazhu Fu. Hallucination detection and correction in medical vlms via counter-evidence verification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, volume 16878 of *Lecture Notes in Computer Science*. Springer Nature Switzerland, 2026.

A Grounded Drift Detection Details

Verifier architecture and training. The BioViL-T encoders remain frozen. For spatial vectors $\mathbf{p}_k \in \mathbf{P}_x$, claim-conditioned attention is

$$\alpha_k(c) = \frac{\exp(\mathbf{p}_k^\top \mathbf{z}_c)}{\sum_r \exp(\mathbf{p}_r^\top \mathbf{z}_c)}, \qquad \mathbf{r}_c = \sum_k \alpha_k(c)\mathbf{p}_k. \tag{12}$$

Table 5: **Lookahead-horizon study on MAIRA-2.** Cov.@90P is coverage when selected-repair precision is constrained to at least 90%.

Method / H	Pref. Acc.	Cov.@90P	CREPE↓	RadG↑	Lat.×
Local PRM	70.8	6.1	–	–	1.04
$H = 0$	74.6	10.2	1.39	35.2	1.09
$H = 1$	79.3	16.4	1.34	35.6	1.16
$H = 2$	**82.9**	**21.0**	**1.26**	**36.2**	1.22
$H = 3$	82.2	20.4	1.28	36.0	1.30

The verifier feature is

$$\mathbf{h}_c = [\mathbf{z}_x; \mathbf{z}_c; \mathbf{z}_x \odot \mathbf{z}_c; |\mathbf{z}_x - \mathbf{z}_c|; \mathbf{r}_c; \mathbf{r}_c \odot \mathbf{z}_c] \in \mathbb{R}^{768}, \tag{13}$$

followed by a $768 \to 256 \to 1$ MLP with GELU and dropout 0.1. Only this head is optimized. We use 2.40M balanced training pairs (1.20M real, 1.20M synthetic), 120k validation pairs, and 120k held-out test pairs after patient-level filtering. Synthetic examples are balanced across finding substitution, polarity reversal, and anatomical relocation. Optimization uses AdamW, learning rate 10^{-4}, weight decay 10^{-2}, batch size 512, three epochs, bfloat16, and early stopping on validation AUPRC. At the selected operating point, held-out support-verification precision/recall/F1 is 91.4/87.8/89.5 for findings, 90.2/82.9/86.4 for polarity, and 89.3/74.6/81.3 for location; localization mIoU is 0.48.

Repair rules. Wrong polarity is flipped only when the opposite assertion is supported. Unsupported positive findings can be removed. Unsupported locations are removed when the location-free claim remains supported; replacement is allowed only if one alternative location is uniquely supported. Unsupported detachable attributes may be deleted but are never synthesized. Measurements, device depth, temporal comparison, and unsupported severity are not introduced. After realization, RadGraph-XL must confirm that every non-target assertion, certainty state, and location relation is unchanged.

Dataset cohorts. The controlled MIMIC-CXR AP/PA Findings cohort contains 162,969 training, 1,286 validation, and 2,461 test studies. The fixed IU-Xray 7:1:2 split contains 2,769/395/791 studies. Preference learning and verifier training are patient-disjoint from the final MIMIC validation/test patients. IU-Xray is used without preference-model update or threshold recalibration.

B Counterfactual Data and Preference Model

Offline pair construction. We sample 16,700 patient-disjoint MIMIC-CXR training studies and generate one report with each of CheXagent, LLaVA-Rad, and MAIRA-2, yielding 50,100 initial report instances. GDD identifies 24,386 parser-level discrepancy states; 13,642 survive visual verification and repair-validity checks and are expanded into corrected/unchanged complete branches. Conservative GREEN dominance retains 8,714 pairs and discards 4,928 ties or error–coverage tradeoffs. Among retained pairs, 5,394 (61.9%) prefer correction and 3,320 (38.1%) prefer leaving the original state unchanged. This non-trivial $U \succ C$ fraction is why a grounded proposal is not treated as an automatic instruction to edit.

Generated branches are cached and then scanned for later proposals. Consequently, a later training state may contain earlier repairs in its prefix without requiring a pretrained online selector. For each individual pair, later interventions remain disabled while its two complete outcomes are generated.

Preference-model optimization. We fine-tune Qwen2.5-0.5B plus a scalar head with AdamW for three epochs, learning rate 10^{-5}, weight decay 0.01, linear warm-up ratio 0.05 and decay, maximum sequence length 1024, gradient norm 1.0, bfloat16, and gradient checkpointing. Effective batch size is 16 pairs. Seeds are $\{17, 23, 42\}$. Minibatches first sample generator identity uniformly and then sample a pair from that generator, preventing high-yield generators from dominating training.

Threshold calibration. The main target is $\pi_{\min} = 0.90$. The validation-selected margin is $\gamma^* = 0.37$, with 90.2% validation precision and 21.0% coverage; on the held-out MAIRA-2 test

Algorithm 1 Online Repair Selection with Lookahead Reuse

Require: x, c, G, grounded detector D, repair operator $\mathcal{R}$, scorer f_θ, horizon H, threshold γ^*

$y \leftarrow \emptyset$, buffer $\mathcal{B} \leftarrow \emptyset$
while EOS not reached **do**
  if $\mathcal{B} \neq \emptyset$ **then**
    $u_t \leftarrow \mathrm{PopFront}(\mathcal{B})$
  else
    $u_t \leftarrow G(x, c, y)$
  end if
  $y_{\leq t} \leftarrow y \oplus u_t$; $\rho_t \leftarrow D(x, y_{\leq t})$
  if $\rho_t = \varnothing$ **then**
    $y \leftarrow y_{\leq t}$; **continue**
  end if
  $\mathcal{B} \leftarrow \emptyset$
  $y^U_{\leq t} \leftarrow y_{\leq t}$; $y^C_{\leq t} \leftarrow \mathcal{R}(y_{\leq t}; \rho_t)$
  $L^U_{t,H} \leftarrow G_H(x, c, y^U_{\leq t})$; $L^C_{t,H} \leftarrow G_H(x, c, y^C_{\leq t})$
  $S^U \leftarrow y^U_{\leq t} \oplus L^U_{t,H}$; $S^C \leftarrow y^C_{\leq t} \oplus L^C_{t,H}$
  $m_t \leftarrow f_\theta(c, S^C) - f_\theta(c, S^U)$
  if $m_t \geq \gamma^*$ **then**
    $y \leftarrow y^C_{\leq t}$; $\mathcal{B} \leftarrow L^C_{t,H}$
  else
    $y \leftarrow y^U_{\leq t}$; $\mathcal{B} \leftarrow L^U_{t,H}$
  end if
end while
return y

experiment the realized Apply-P is 91.2% with 21.5% intervention rate (Table 4). Varying $\pi_{\min}$ from 0.80 to 0.95 gives the expected precision–coverage trade-off; the threshold is therefore primarily a reliability control rather than the source of the full quality gain.

C Additional Experiments and Analysis

Independent agreement with the offline supervisor. On held-out corrected/unchanged pairs retained by conservative GREEN dominance, the GREEN ordering agrees with the independent CREPE, RaTEScore, and RadGraph orderings in 78.6%, 74.9%, and 72.3% of pairs, respectively; all three agree on 64.1%. These statistics are diagnostic only and are never used to train or calibrate f_θ.

Runtime and intervention frequency. On H800 GPUs, average end-to-end report latency changes from 3.8 to 4.7 s for CheXagent (1.24×), 3.2 to 4.1 s for LLaVA-Rad (1.28×), and 4.5 to 5.5 s for MAIRA-2 (1.22×). Average extra generated tokens are 18–23% above base generation, depending on the generator. Counterfactual report collection dominates offline cost; at inference, the selected lookahead is reused rather than regenerated.

Failure modes. The remaining errors fall into four practical categories: ambiguous or incorrect visual grounding; over-specific anatomical localization; preference-selection errors in which a locally valid repair receives the wrong branch ordering; and missed actionable errors caused by parsing failures or content outside the supported repair taxonomy. These cases define the boundary of the current intervention space and motivate conservative abstention rather than broader free-form rewriting.