Source-Causal Control of Historical Context in Longitudinal Radiology Report Generation

Dmitry Lvov[1,2] **Ilya Pershin**[1,3]
[1]Research Center of the Artificial Intelligence Institute,
Innopolis University, Innopolis, Russia
[2]ITMO University, Saint Petersburg, Russia
[3]Kazan Federal University, Kazan, Russia
{d.lvov,i.pershin}@innopolis.ru

Abstract

Historical context helps radiology report generators describe change over time, but it can also introduce findings whose support in the current examination is unclear. We introduce source-causal robustness, an audit that holds the current chest X-ray fixed while intervening on historical images and reports. Across 18,716 MIMIC-CXR pairs and 93,580 generations, matched wrong-history increases reference-derived false positives by .366 per report while reducing false negatives by .129. We develop the Source-Causal Finding Router (SCFR), a text-only controller that selects among source-valid reports and finding-filtered proposals, using the current-only generated report as its anchor. On three 300-study controller subsets, SCFR reduces matched-wrong-history proxy false positives by .343–.550 per report, passes Holm-adjusted testing ($p \leq .0012$), and meets fixed current-only point-estimate quality bounds. Blinded radiologist assessment of 60 examinations calibrates the text proxies against current-image support; a separate 30-case study examines selected history-conditioned content. Additional generators and structured audits characterize the resulting source sensitivity and routing tradeoffs. The framework provides a reusable historical-source audit and a lightweight controller for frozen report generators. Code and reproduction instructions: `https://github.com/LvovDmitro/SCFR`.

1 Introduction

A chest X-ray report describes a patient at a particular point in time. Previous examinations help a radiologist recognize a new opacity, follow an effusion, or distinguish a persistent abnormality from a recent change. Report generators increasingly receive the same longitudinal context through prior images and reports [1–4]. This creates a source-selection problem: the report must describe the current examination while drawing useful comparisons with the past. A finding stated in a previous report can remain linguistically plausible even after it has resolved, and an unrelated history can introduce an equally plausible finding. Figure 1 illustrates how prior text can change the reported finding while the current image stays fixed.

The difficulty becomes apparent when evaluating these systems. A history-conditioned report may recover more findings mentioned in a reference report while also introducing additional positives. Aggregate quality scores can obscure this tradeoff, particularly when common findings and familiar report language receive credit. Text-overlap metrics, clinical labelers, and structured factuality metrics provide complementary measures of report quality [5–12]. To understand historical carryover, evaluation also needs to establish how the same current examination is reported under different historical sources.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

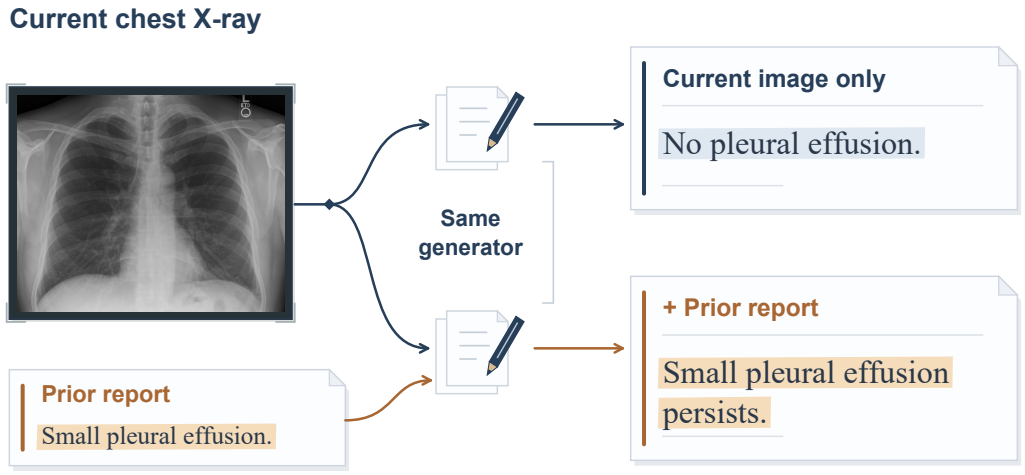

Figure 1: Source-dependent reporting. Both paths use the same current image and generator; prior-report text enters only the lower path, where the historical finding reappears. Report excerpts are authored schematic examples; the public radiograph supplies image context.

We study this question through *source-causal robustness*: holding the current chest X-ray fixed while intervening on the historical input. Alongside valid history, we supply history from another subject, either randomly selected or matched by finding signature and prior-report length. The matched wrong-history arm tests a challenging failure mode in which the context is plausible but belongs to the wrong patient. Paired changes in extracted findings quantify the generator's response to this intervention. A separate blinded image review determines whether sampled additions are supported by the current examination, allowing legitimate history-aware interpretation to be distinguished from image-unsupported content.

The source-selection problem also calls for a way to control historical influence after generation. We develop the Source-Causal Finding Router (SCFR), which uses the current-only generated report to anchor selection among source-valid reports and finding-filtered proposals. The study makes three contributions:

- A source-intervention audit that measures report changes under valid, random wrong, and finding-matched wrong history while holding the current examination fixed.
- A lightweight text-only router for frozen generators that selects among current-only and valid-history candidates, returning the selected candidate unchanged.
- An aligned evaluation on MIMIC-CXR that separates wrong-history stress effects from current-only quality comparisons, with blinded radiologist assessment and additional generator and structured-label audits.

2 Related Work

Radiology report generation. Early neural systems generated reports from image features using attention and cross-modal representations, followed by memory-based approaches for long-form clinical text [13–15]. MIMIC-CXR and IU X-Ray enabled the development and comparison of these methods [16, 17]. More recent systems use frozen language models, clinical instruction tuning, and conversational supervision, including R2GenGPT and RaDialog [18, 19]. LLaVA-Med, Med-Flamingo, BioViL, CheXagent, and MedGemma extend the range of biomedical vision-language capabilities [20–23, 4]. Temporal modeling brings prior studies directly into this setting: BioViL-T learns from temporal structure, while MAIRA-2, Libra, and MAIRA-X incorporate longitudinal information into reporting [24, 1–3]. As this input becomes richer, evaluation must account for the relationship between the historical source and the current patient.

Clinical factuality and temporal evaluation. BLEU, ROUGE, and BERTScore measure textual correspondence, while CheXpert and CheXbert support comparisons at the level of extracted clinical findings [5–9]. RadGraph represents entities and relations, and GREEN and RadFact assess factual discrepancies with structured or language-model-assisted reasoning [10–12]. Resources addressing temporal language and localized findings provide further resolution: RadGraph2 tracks temporal finding relations, ReXPref-Prior examines prior-examination language, ReXErr supplies clinically meaningful report errors, and Chest ImaGenome associates findings with anatomical regions [25–28]. These measures characterize report content, but a reference-discordant finding may still be supported by the current image. Source dependence and current-image support therefore require separate assessment.

Grounding and source interventions. Attributing a statement to an input is a separate problem from measuring its textual quality. Grad-CAM provides image-localization cues, while work on attention as explanation cautions against treating internal attention weights as sufficient evidence of grounding [29, 30]. Invariant prediction, invariant risk minimization, and counterfactual explanations offer related ways to reason about changes in inputs and environments [31–33]. For longitudinal reporting, this perspective raises a source-specific question: how much does a report depend on the identity of the supplied history when the current examination is unchanged?

Controlling report errors. Factuality can also be addressed during decoding, candidate selection, or model training. Clinical Contrastive Decoding incorporates structured expert-image signals into training-free token-level decoding [34]. Sentence-level verification and weighted best-of-N selection provide another route to choosing reports [35]. Direct preference optimization has been used to suppress hallucinated references to prior examinations [36]. These approaches address report errors at different stages of generation. A remaining evaluation need is to distinguish improved agreement with a reference from sensitivity to plausible history belonging to another patient.

3 Method

3.1 Auditing historical sources

Let x_i be the current chest X-ray for study i, r_i its reference report, and G a fixed report generator. We evaluate five historical-source conditions: current-only (cur), the patient's real prior images and report (real), the real prior report alone (text), random wrong-history (rand), and matched wrong-history (match). For source intervention $\mathcal{I}_a$, the generated report is

$$\hat{y}_i^a = G(x_i, \mathcal{I}_a(i)), \qquad a \in \{\mathrm{cur}, \mathrm{real}, \mathrm{text}, \mathrm{rand}, \mathrm{match}\}. \tag{1}$$

The current image is identical across arms. Real-history arms retain the patient's source alignment; wrong-history arms replace historical images and text with another subject's context.

Matched donors belong to a different subject and satisfy the target's split restriction. The rule first minimizes the difference between finding signatures extracted from target and donor current reference reports, then the difference in prior-report word count; seeded tie-breaking completes the assignment. Prior images and text move together, and donors may be reused. This reference-conditioned matching is an offline audit procedure (Appendix H).

For a metric m and an evaluation population $\mathcal{D}$, we report the paired source effect

$$\Delta_m(a, b) = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \big[m(\hat{y}_i^a, r_i) - m(\hat{y}_i^b, r_i) \big]. \tag{2}$$

Comparing a historical arm with current-only measures its source-induced change. Comparing SCFR with matched wrong-history places the controller relative to this stress condition; direct current-only comparisons assess quality separately.

3.2 Finding-level measurement

We extract a fixed ledger of 11 finding categories, recording positive, negative, uncertain, and absent states with local negation handling. Let $P(y)$ contain the categories asserted positively in report y.

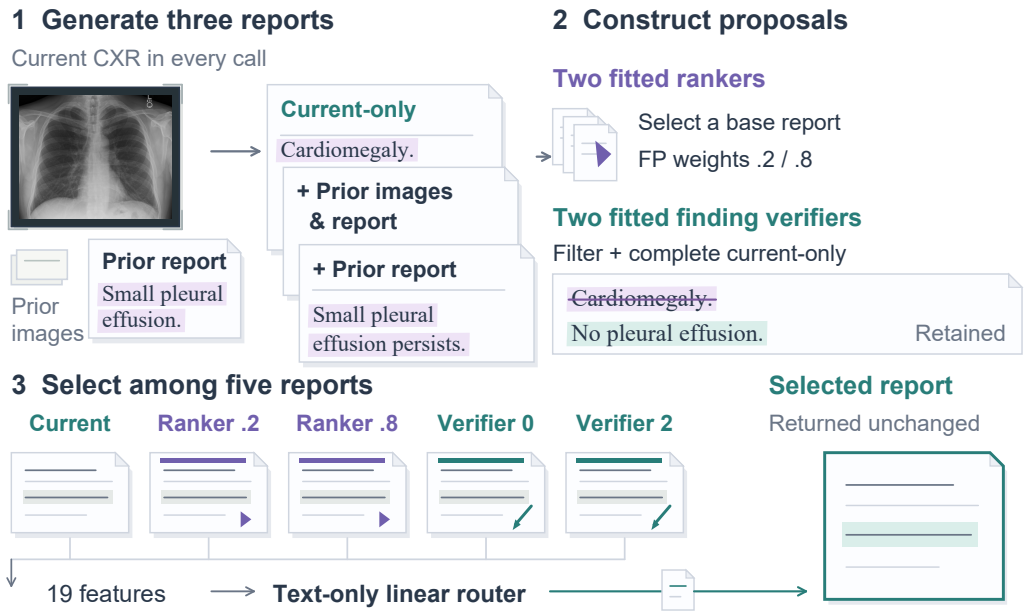

Figure 2: SCFR deployment. Three calls share the current image. Two fitted rankers select base reports; two independently fitted verifiers edit current-only text. These four proposals join the unchanged current-only anchor. A text-only linear router uses 14 shared text features and five candidate indicators to select a report unchanged. Ranker labels denote FP weights; verifier labels denote nominal relative-length allowances (Appendix A). Authored snippets illustrate sentence deletion and retention; the public radiograph supplies image context.

The automatic error counts are

$$\begin{aligned} \mathrm{FP}(y, r) &= |P(y) \setminus P(r)|, \\ \mathrm{FN}(y, r) &= |P(r) \setminus P(y)|. \end{aligned} \tag{3}$$

These *proxy FP* and *proxy FN* count reference-discordant positives and missed reference positives. Finding F1 and status accuracy summarize correspondence with the same ledger. A supported statement may be absent from a reference report, so the reader study assesses current-image support separately (Section 4.4). Appendix A defines the extractor and categories.

3.3 Constructing source-valid candidates

SCFR uses three base generations: current-only y_i^{cur}, real-history y_i^{real}, and prior-report-only y_i^{text}. Figure 2 follows these reports through proposal construction and final text-only selection.

Four proposals are constructed from these reports:

1. Two frozen proposal rankers select a base report unchanged. Their fitting targets are finding F1 minus weighted proxy errors, with FP weights .2 and .8 and a shared FN weight .1, yielding $b_i^{.2}$ and $b_i^{.8}$.
2. Two independently fitted logistic finding verifiers predict allowed finding sets from the three base reports. Each removes current-only sentences containing disallowed positives and appends compact canonical statements for missing allowed findings. The proposals b_i^{V0} and b_i^{V2} have nominal append budgets of zero and two tokens relative to current-only. An empty-text fallback can exceed these budgets (Appendix A).

The final candidate order is

$$\mathcal{B}_i = \left(y_i^{\mathrm{cur}}, b_i^{.2}, b_i^{.8}, b_i^{V0}, b_i^{V2}\right). \tag{4}$$

This retains the unchanged current-only anchor alongside base-report selections and finding-filtered text. Proposal construction is text-only; the complete system uses three generator calls per study.

3.4 Learning and applying the router

The final router is an averaged structured perceptron with 19 features $\phi(i, b)$ per candidate: 14 shared text features and five candidate indicators.

Feature family	Count	Contents
Report form	4	Bias, log length, positive-finding count, temporal-term count
Text agreement	5	Prior trigram-copy precision; sequence similarity to valid prior text and each of the three base reports
Finding differences	5	Positive additions and removals relative to current-only and valid prior text; positive overlap with prior text
Candidate identity	5	One-hot indicators in the order of Equation (4)

Appendix A gives the exact feature order, scaling, and coefficients. During fitting, each candidate receives the reference-derived utility

$$U(i, b) = \mathrm{F1}(b, r_i) - .2\,\mathrm{FP}(b, r_i) - .4\,\mathrm{FN}(b, r_i). \tag{5}$$

The first utility maximizer is the training target. For each evaluated subset E, the *final-ranker* fitting pool excludes every row belonging to a subject in E. The update rule and pool sizes appear in Appendix A; operating-point selection and sensitivity analyses appear in Appendix F.

At deployment, the frozen weights select

$$\hat{y}_i^{\mathrm{SCFR}} = \operatorname{first} \operatorname*{argmax}_{b \in \mathcal{B}_i} w_E^\top \phi(i, b). \tag{6}$$

Candidate order resolves ties, and SCFR returns the selected text unchanged. The router receives generated reports and valid prior text. Its current-study anchor is the current-only *generated report*; it receives neither pixels nor reference reports at deployment. Wrong-history generations are used only in the offline audit and never enter the router or its candidate set.

The stress endpoint is the proxy-FP difference from matched wrong-history, with Holm-adjusted tests. On the same studies, direct comparisons with always-current-only report paired uncertainty and check fixed point-estimate bounds: at most .02 lower finding F1 or status accuracy and at most .10 additional FP or FN per report. These bounds specify the selected quality tolerance.

4 Results

4.1 Evaluation setup

We use MIMIC-CXR-JPG images and MIMIC-CXR reports to construct longitudinal pairs with a current examination, prior imaging and report, and current reference [37, 16]. The full source intervention comprises 18,716 paired studies and five reports per study, giving 93,580 generations. Every arm uses the same current images and reference-based metrics. We report finding F1, finding-status accuracy, and proxy FP/FN per report.

Three 1,000-study parent cohorts provide the source-intervention comparisons. Their frozen 300-study controller subsets contain 230, 231, and 226 subjects in A, B, and C. Final-ranker fitting pools contain 1,466, 1,343, and 1,330 rows after evaluation-subject exclusion. The utility setting was selected using a 400-study pilot and the B/C cohorts. A was generated after selection, using the recorded development-study exclusion list; B/C provide descriptive rechecks of the reused development cohorts. Tables 2 and 3 explicitly separate the parent and controller populations.

Controller-comparison uncertainty uses 5,000 recipient-subject cluster bootstrap resamples, conditional on the frozen report outputs and donor assignments. The three primary SCFR-minus-matched-history FP tests use Holm adjustment. Direct-current comparisons use paired intervals on the same controller subsets. The fixed quality tolerances in Section 3 apply to point estimates. Blinded image assessment and structured audits complement these reference-derived endpoints [25–28].

4.2 Historical-context intervention

Historical context changes the balance between reference coverage and additional positive findings (Table 1). Real history raises finding F1 from .4338 to .4375 and lowers FN by .1107 per report, while

increasing FP by .2607. Matched wrong-history also lowers FN, by .1290, but raises FP by .3662, a 49.4% increase over current-only. Random wrong-history produces the largest FP increase, .4646 per report. Thus an apparent recall benefit can accompany a substantial increase in reference-discordant findings even when the supplied history belongs to another subject.

Table 1: Source conditions on the same 18,716 studies. All finding metrics are reference-derived. Bold and italics mark the numerical best and second-best values per metric; lower FN alone does not establish valid use of history.

Source condition	Finding F1 ↑	Status ↑	FP/report ↓	FN/report ↓
Current-only	*.4338*	**.7055**	**.7416**	1.2122
Real prior	**.4375**	*.6954*	*1.0023*	*1.1015*
Prior report only	.4095	.6868	1.0320	1.1624
Random wrong-history	.3883	.6744	1.2062	1.1441
Matched wrong-history	.4252	.6874	1.1078	**1.0832**

The complete parent cohorts show the same direction for the primary generator (Table 2). Real and matched wrong-history both increase FP while reducing FN. This agreement across cohorts motivates evaluating source sensitivity jointly with reporting quality, rather than treating added context as an unconditional improvement.

Table 2: Complete parent cohorts, with studies/subjects in parentheses. All deltas are relative to current-only. RP denotes real prior; WH denotes matched wrong-history. Bold identifies the primary WH FP stress endpoint.

Cohort	RP ΔFP	RP ΔFN	WH ΔFP	WH ΔFN	WH ΔStatus
A (1000/750)	+.278	-.101	**+.348**	-.135	-.015
B (1000/736)	+.323	-.087	**+.454**	-.111	-.027
C (1000/749)	+.361	-.133	**+.426**	-.171	-.018

4.3 Source-valid report routing

Table 3 compares the current-only output, matched-history stress output, and SCFR selection on identical controller populations. SCFR uses source-valid candidate reports, while WH supplies the separate stress reference. The controller removes 103, 165, and 137 proxy FP relative to WH. Dividing these reductions by the corresponding WH excess over current-only gives **108.4%, 92.7%, and 95.1%** repair. The first value exceeds 100% because SCFR's FP count is eight below current-only in A. All three Holm-adjusted FP tests have $p \leq .0012$.

Table 3: Aligned 300-study controller subsets. Cells give counts and per-report rates. Both difference columns use the same rows and outputs. The WH FP reduction is bold; the direct-current difference is italic. A evaluates the frozen selection; B/C reuse development cohorts.

Panel	Metric	Current	WH	SCFR	SCFR−WH	SCFR−Current
A (300/230)	FP	223 (.7433)	318 (1.0600)	215 (.7167)	**−.3433**	*−.0267*
	FN	372 (1.2400)	321 (1.0700)	372 (1.2400)	+.1700	*.0000*
B (300/231)	FP	225 (.7500)	403 (1.3433)	238 (.7933)	**−.5500**	*+.0433*
	FN	372 (1.2400)	325 (1.0833)	365 (1.2167)	+.1333	*−.0233*
C (300/226)	FP	204 (.6800)	348 (1.1600)	211 (.7033)	**−.4567**	*+.0233*
	FN	373 (1.2433)	331 (1.1033)	376 (1.2533)	+.1500	*+.0100*

The resulting operating point is close to current-only (Figure 3). Direct-current F1 changes are +.0027, +.0085, and -.0097, and FN changes are .0000, -.0233, and +.0100. All fixed point-estimate quality bounds are met; paired intervals quantify the uncertainty of each change. Relative to WH, the FP reduction accompanies FN increases of .1700, .1333, and .1500 per report, returning toward the current-only balance. SCFR selects 139 text-distinct non-current reports across the 900 evaluated studies, retaining an explicit route for using history instead of committing every case to current-only.

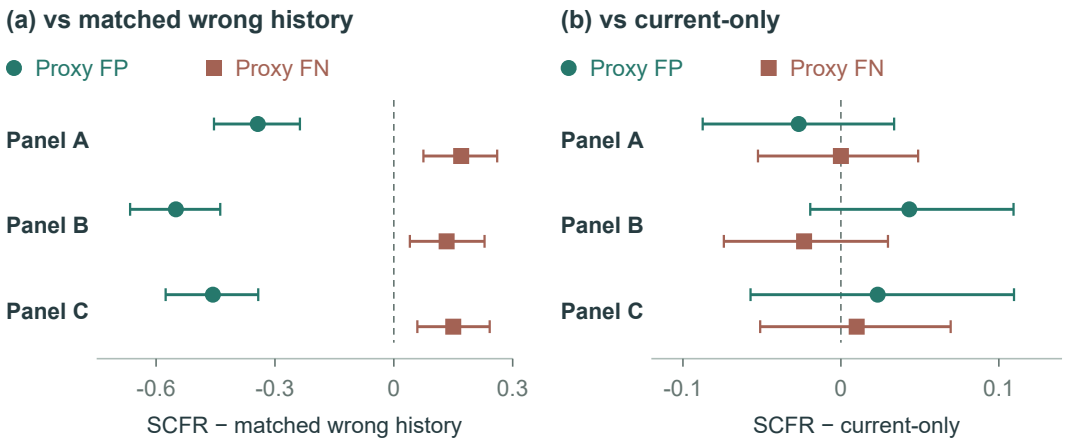

Figure 3: Paired controller effects on the aligned 300-study populations. Points are per-report differences; bars are subject-cluster 95% intervals. The WH comparison measures displacement from the stress condition; the current-only comparison shows the operating point against the one-generation anchor. A is post-selection; B/C are reused development cohorts.

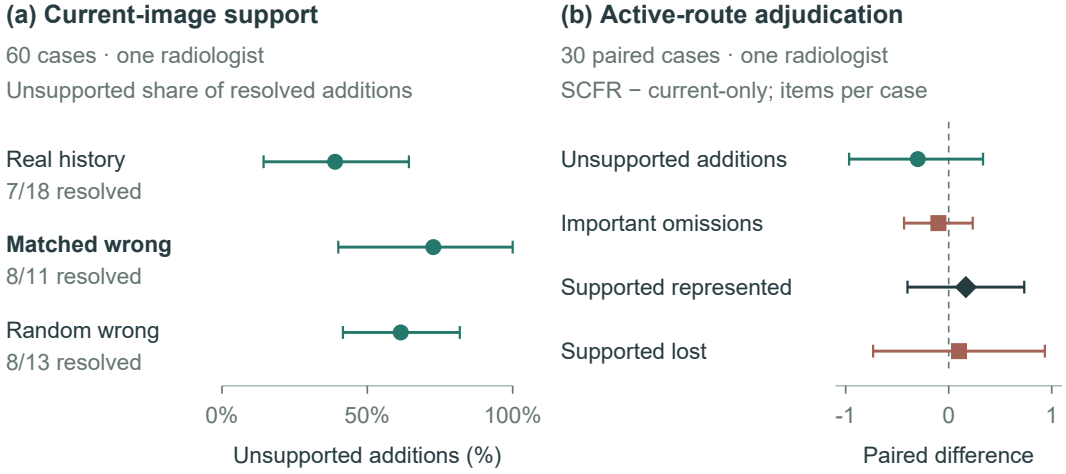

Figure 4: Blinded reader evidence. The calibration panel shows image-unsupported fractions among resolved additions, with 95% intervals and explicit denominators. The active-route panel shows paired item differences between SCFR and current-only on 30 sampled routes. The two tasks assess proxy calibration and selected-report content, respectively.

4.4 Image-grounded assessment

A qualified radiologist assessed 60 prespecified, subject-unique current examinations while blinded to generated reports, historical arms, proxy labels, and routing. In this challenge-stratified sample, agreement between the reference-derived labels and the radiologist was 84.1% over 643 resolved finding decisions ($\kappa = .281$). Image review separates a reference-discordant addition from a finding supported by the current examination.

Among resolved positive additions over current-only, the reader judged 7/18 unsupported under real history and 8/11 under matched wrong-history. The resulting 33.8-percentage-point contrast has a 95% interval of 5.6–63.6 (Figure 4). The valid-history additions also include 11 image-supported findings, compared with three for matched history. Appendix C reports uncertain additions, strata, inclusion weights, and complete instructions; weighted estimates target the defined 5,138-subject challenge frame.

A separate blinded comparison examined 30 cases where SCFR selected text distinct from current-only. The reader found true history informative in 25/30 cases. SCFR represented 156 supported findings and lost 50, versus 151 and 47 for current-only. Unsupported additions totalled 17 versus 26, and important omissions 17 versus 20. These paired differences and their intervals are displayed together; the preference counts were 13 for SCFR and 17 for current-only (exact sign-test $p = .585$). SCFR expressed four correct and one partially correct temporal relation, demonstrating the type of longitudinal content retained on these active routes.

4.5 Additional generators and structured audits

The intervention reveals generator-specific responses (Table 4). Whereas MAIRA-2 trades lower FN for higher FP under matched wrong-history, RadInfer lowers both proxy counts and Libra lowers FP while increasing FN. Applying the same intervention across backbones therefore exposes different uses of historical context.

Table 4: Source-intervention panels across generators. WH deltas are relative to current-only, with current images fixed within each panel. Each row describes that generator's population.

Generator	Studies	Current finding F1	WH ΔFP	WH ΔFN
MAIRA-2	18,716	.4338	+.3662	-.1290
RadInfer	400	.1450	-.0950	-.2175
Libra	306	.3851	-.1830	+.3824

We also evaluate controller transfer with five subject-disjoint folds per target generator. Proposal models are fitted within the target folds, and the final 19-feature router remains frozen. A separate meta-fold checks the four current-only quality bounds before deployment on the test fold. RadInfer improves finding F1 by **.1060** (95% CI .0634–.1460), reduces FP by .2925, and reduces FN by .1200 per report. Libra reduces FP by .0719 while increasing FN by .0752; three of its five folds use current-only fallback. Appendix E gives all counts, folds, intervals, and gate decisions.

Structured audits examine temporal language and anatomical labels (Table 5). Prior-cue rates decrease across all three controller evaluations. ReXPref-Prior records a reduction from 14.6% to .8% in unsupported prior-examination cues. RadGraph2 records 93 WH FP findings removed in its 130-study audit, and Chest ImaGenome provides an anatomy-linked assessment with FP delta -.3615 and status delta +.0329.

Table 5: Structured report audits. Before/after values follow each audit's own criteria and population. Bold emphasizes the SCFR outcome.

Audit	WH / before	SCFR / after	Additional result
Prior-cue A	8.7%	**1.0%**	88.5% cue reduction
Prior-cue B	10.3%	**5.0%**	51.5% cue reduction
Prior-cue C	10.7%	**2.0%**	81.3% cue reduction
ReXPref-Prior	14.6%	**.8%**	94.5% cue reduction
RadGraph2	130 studies	**93 FP removed**	27/55 added findings current/reference-supported
Chest ImaGenome	—	**FP** $-.3615$	Status $+.0329$; FP $p = .0012$

The lexical decomposition further separates historical influence from literal copying. Matched wrong-history introduces 18,451 positive additions over current-only. Of these, 40.8% name a finding also present in the supplied prior report. Exact sentence matches account for .5%; near matches, including exact matches, for 2.2%. Same-prior-finding additions below the near-copy threshold account for 38.8%. Among the 11,969 reference-discordant additions, 31.8% share a prior finding and 2.0% are near copies (Appendix G). Historical influence therefore extends beyond repeated strings, motivating checks at the finding level alongside lexical overlap.

4.6 Filtering baselines

Simple text filters provide a useful reference for the precision–recall balance. On the complete 18,716-study snapshot, prohibiting every positive addition reduces FP from .7416 to .6146, while

increasing FN from 1.2122 to 1.2922 (Table 6). Its finding-F1 difference from current-only is $-.0071$ (95% CI $[-.0093, -.0050]$); the FP and FN differences are $-.1270$ and $+.0800$, respectively. A rule that blocks prior-only additions has higher FP and FN than current-only. The benefit of suppressing an addition therefore depends on what that addition contributes to the report.

Table 6: Filtering controls on the complete 18,716-study snapshot, separate from the 300-study controller cohorts. Bold and italics identify the numerical best and second-best deployable controls per metric. The outcome-conditioned oracle is an offline diagnostic reference.

Control	Finding F1 ↑	Status ↑	FP ↓	FN ↓
Always current-only	**.4338**	*.7055*	*.7416*	**1.2122**
No positive additions	*.4267*	**.7073**	**.6146**	1.2922
Block prior-only additions	.4109	.6963	.7931	*1.2541*
Outcome-conditioned oracle	.5345	.7361	.6750	.9603

These controls clarify the role of SCFR's candidate construction. A current-only anchor supplies the unfiltered current description. Precision-oriented proposals remove positive content more aggressively, while recall-oriented proposals can retain history-conditioned content. The final ranker chooses among these alternatives under a common reference-derived fitting objective. This separates the audit of historical influence from the policy used to admit or suppress its report content.

4.7 Components and operating-point selection

The component comparison shows why the candidate set contains different precision–recall preferences (Figure 5). The recall-biased proposal has the highest finding F1 and lowest FN, accompanied by higher FP. Verifier proposals favor lower FP and higher status accuracy. The current-only anchor and SCFR occupy an intermediate region. SCFR chooses among these alternatives for each examination, meeting the panel-specific point-estimate quality bounds. Appendix F reports all four metrics in Table 16.

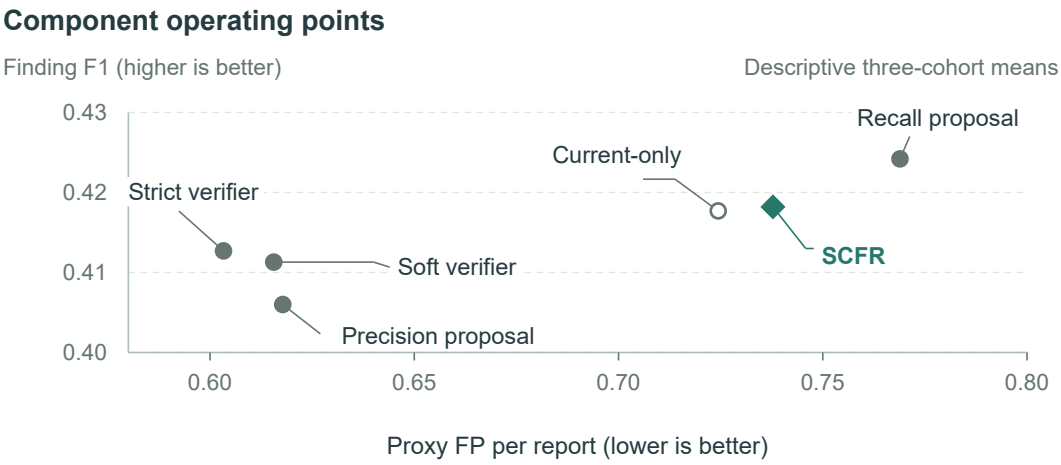

Figure 5: Component tradeoffs. Higher finding F1 and lower FP are preferred. Each point is a descriptive unweighted mean over the three controller cohorts. Table 16 gives the complete values, including status accuracy and FN.

Table 7 shows how the utility weights select the balance before the post-selection A evaluation. Increasing the FP penalty reduces FP more aggressively but also increases FN. The selected pair $(.2, .4)$ has the highest mean finding F1 among the displayed settings and reduces both proxy counts on the development panels. This criterion fixes the preference among available candidates, while the quality bounds determine whether the selected operating point can be deployed.

Sensitivity to 10, 25, 50, and 100 epochs produces between 138 and 142 non-current routes, with mean finding F1 between .4182 and .4194. Halving the quality tolerances to $.01/.01/.05/.05$ preserves the

Table 7: Development utility sweep: unweighted mean SCFR-minus-current-only effects on the pilot and the parent cohorts for B/C. Bold marks the selected setting. These development effects guide selection; the aligned post-selection A results are reported separately in Table 3.

λ_{FP}	λ_{FN}	ΔF1	ΔFP	ΔFN
.2	**.4**	$\mathbf{+.0060}$	$\mathbf{-.0146}$	$\mathbf{-.0187}$
.4	.4	$-.0016$	$-.0420$	$+.0159$
.6	.3	$-.0054$	$-.1040$	$+.0540$

pass decision on all three panels. Appendix F gives the proposal weights, verifier append caps, and full epoch results.

5 Limitations and responsible use

The study is retrospective and evaluates reporting on MIMIC-CXR, primarily with MAIRA-2. The generator has MIMIC-CXR training-domain exposure, with unknown overlap with the audited cases. Subject exclusion applies to final-ranker fitting; upstream generator and historical proposal fits require separate provenance checks. Matching uses current-reference signatures and prior-report length, and bootstrap intervals condition on fixed donor assignments and fitted models. Historical images and text change together, so the intervention measures their joint effect.

Reference-derived FP/FN measure text agreement. Image-grounded calibration uses one reader, a challenge-stratified sample, modest chance-corrected agreement, and small addition denominators. Direct-current recall and F1 effects are mixed; paired active-route intervals include zero and overall preference is inconclusive. The fixed bounds describe a point-estimate operating point. SCFR assumes patient-aligned history and requires three generator calls compared with one for current-only. Prospective multi-reader evaluation should assess clinical omissions, temporal correctness, and automation bias before deployment.

6 Conclusion

Longitudinal radiology report generation requires control over how historical information enters a current description. Holding the current chest X-ray fixed reveals changes that ordinary report-quality comparisons can obscure: additional history can recover reference findings while also introducing positives that disagree with the current reference. The matched wrong-history intervention makes this dependence observable under plausible but patient-mismatched context. Blinded image assessment then connects the text audit to current-image support, distinguishing supported additions from image-unsupported content.

SCFR translates this evaluation into a report-selection policy for frozen generators. Three source-valid generations supply the current-only anchor, base-report alternatives, and finding-filtered proposals. A 19-feature text ranker selects one complete report unchanged. Across the aligned controller cohorts, this selection removes 108.4%, 92.7%, and 95.1% of the WH proxy-FP excess while meeting the fixed current-only quality bounds. The 139 non-current selections among 900 studies preserve an explicit channel for history-conditioned reporting at this operating point. The active-route study further identifies the content affected by these decisions, including supported findings, omissions, and temporal relations.

The cross-generator evaluations show why historical-source control should be assessed for each backbone. RadInfer supports active routing across all five evaluation folds, whereas Libra uses current-only fallback in three folds. Together with the component and utility analyses, these results expose how candidate selection trades coverage for fewer positives and when the resulting policy meets the current-only quality criteria. Generator recall changes, image support for sampled additions, and controller effects relative to current-only are assessed within their respective evaluation populations, providing complementary views of historical-source dependence. The audit measures changes induced by the supplied source; the controller determines which available report is returned.

Acknowledgments and Disclosure of Funding

The study was supported by the Ministry of Economic Development of the Russian Federation (agreement No. 139-10-2025-034 dated 19 June 2025, IGK 000000C313925P4D0002). The authors declare no competing interests.

References

[1] Shruthi Bannur, Kenza Bouzid, Daniel Coelho de Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024. URL `https://arxiv.org/abs/2406.04449v1`.

[2] Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. Libra: Leveraging temporal images for biomedical radiology analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17275–17303, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.888. URL `https://aclanthology.org/2025.findings-acl.888/`.

[3] Harshita Sharma, Maxwell C. Reynolds, Valentina Salvatelli, Anne-Marie G. Sykes, Kelly K. Horst, Anton Schwaighofer, Maximilian Ilse, Olesya Melnichenko, Sam Bond-Taylor, Fernando Pérez-García, Vamshi K. Mugu, Alex Chan, Ceylan Colak, Shelby A. Swartz, Motassem B. Nashawaty, Austin J. Gonzalez, Heather A. Ouellette, Selnur B. Erdal, Beth A. Schueler, Maria T. Wetscherek, Noel Codella, Mohit Jain, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Stephanie Hyland, Panos Korfiatis, Ashish Khandelwal, and Javier Alvarez-Valle. Closing the performance gap between ai and radiologists in chest x-ray reporting. *arXiv preprint arXiv:2511.21735*, 2025. URL `https://arxiv.org/abs/2511.21735`.

[4] Google Health AI Developer Foundations. MedGemma. `https://developers.google.com/health-ai-developer-foundations/medgemma`, 2026. Model documentation, last updated 2026-01-13.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *ACL*, 2002.

[6] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020.

[8] 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 Halabi, Jesse 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 *AAAI*, 2019.

[9] 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 *EMNLP*, pages 1500–1519, 2020. doi: 10.18653/v1/2020.emnlp-main.117.

[10] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH 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. *arXiv preprint arXiv:2106.14463*, 2021. URL `https://arxiv.org/abs/2106.14463`.

[11] 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*, pages 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/.

[12] Microsoft Research. RadFact: An llm-based evaluation metric for ai-generated radiology reporting. https://github.com/microsoft/RadFact, 2024.

[13] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. TieNet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *CVPR*, 2018.

[14] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *EMNLP*, 2020.

[15] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *ACL-IJCNLP*, 2021.

[16] 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 large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. URL https://arxiv.org/abs/1901.07042v1.

[17] 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.

[18] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology*, 1(3):100033, 2023. doi: 10.1016/j.metrad.2023.100033.

[19] Chantal Pellegrini et al. RaDialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*, 2023. URL https://arxiv.org/abs/2311.18681.

[20] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *NeurIPS Datasets and Benchmarks Track*, 2023.

[21] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo P. Reis, and Pranav Rajpurkar. Med-Flamingo: A multimodal medical few-shot learner. In *Machine Learning for Health*, 2023.

[22] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. Making the most of text semantics to improve biomedical vision-language processing. In *ECCV*, 2022.

[23] Zhihong Chen et al. CheXagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024. URL https://arxiv.org/abs/2401.12208v1.

[24] 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 *CVPR*, pages 15016–15027, 2023.

[25] Adam Dejl, Sameer Khanna, Patricia Therese Pile, Kibo Yoon, Steven QH Truong, Hanh Duong, Agustina Saenz, and Pranav Rajpurkar. RadGraph2: Tracking findings over time in radiology reports. *PhysioNet*, August 2024. doi: 10.13026/q65y-9688. URL https://doi.org/10.13026/q65y-9688. Version 1.0.0.

[26] Oishi Banerjee, Hong-Yu Zhou, Subathra Adithan, Stephen Kwak, Kay Wu, and Pranav Rajpurkar. ReXPref-Prior: A mimic-cxr preference dataset for reducing hallucinated prior exams in radiology report generation. *PhysioNet*, August 2024. doi: 10.13026/t13x-4r94. URL https://doi.org/10.13026/t13x-4r94. Version 1.0.0.

[27] Vishwanatha Rao, Serena Zhang, Julian Acosta, Subathra Adithan, and Pranav Rajpurkar. ReXErr-v1: Clinically meaningful chest x-ray report errors derived from mimic-cxr. *PhysioNet*, March 2025. doi: 10.13026/9dns-vd94. URL https://doi.org/10.13026/9dns-vd94. Version 1.0.0.

[28] Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, Tanveer Syeda-Mahmood, and Mehdi Moradi. Chest ImaGenome Dataset. *PhysioNet*, July 2021. doi: 10.13026/wv01-y230. URL https://doi.org/10.13026/wv01-y230. Version 1.0.0.

[29] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[30] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *NAACL*, 2019.

[31] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B*, 78(5):947–1012, 2016.

[32] Martin Arjovsky, Leon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

[33] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887, 2018.

[34] 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*, pages 35174–35200, 2026. doi: 10.18653/v1/2026.findings-acl.1755. URL https://aclanthology.org/2026.findings-acl.1755/.

[35] 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. URL https://arxiv.org/abs/2510.23217v1.

[36] Oishi Banerjee, Hong-Yu Zhou, Kay Wu, Subathra Adithan, Stephen Kwak, and Pranav Rajpurkar. Direct preference optimization for suppressing hallucinated prior exams in radiology report generation. In *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*, 2024. PMLR. URL https://proceedings.mlr.press/v252/banerjee24a.html.

[37] Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR-JPG - chest radiographs with structured labels. *PhysioNet*, March 2024. doi: 10.13026/jsn5-t979. URL https://doi.org/10.13026/jsn5-t979. Version 2.1.0.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

Justification: The abstract and Introduction describe the historical-source audit and source-valid controller. Section 4 distinguishes the evaluation populations and their selection roles.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 covers the study scope, provenance, proxy and reader uncertainty, mixed direct-current effects, inference cost, and prospective validation.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete proof?

   Answer: [N/A]

   Justification: This is an empirical study with no theoretical results requiring proof.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results to the extent that it affects the main claims and conclusions?

   Answer: [No]

   Justification: Methods and appendices specify the protocol. Exact replay also requires historical component fits, split manifests, and frozen generations absent from the release.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results?

   Answer: [No]

   Justification: Code and refitting instructions are public. Credentialed patient data, historical component fits, split manifests, and frozen generations are not distributed.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 3–4 and the appendices give the protocol, evaluation sizes, metrics, component comparisons, and audit details.

7. **Experiment statistical significance**

   Question: Does the paper report error bars or other appropriate information about statistical significance?

   Answer: [Yes]

   Justification: Section 4 and the appendices report paired bootstrap intervals and Holm-adjusted primary FP tests, distinguishing post-selection Panel A from development-reused Panels B/C.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the compute resources needed to reproduce the experiments?

   Answer: [No]

   Justification: Generation counts and three-call inference cost are reported, but per-experiment hardware and runtime are not fully documented.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

   Answer: [Yes]

Justification: Credentialed, de-identified data remained under their access agreements. Expert annotation was institutionally authorized, with a recorded ethics determination (Section 5; Appendix C).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Introduction and Section 5 discuss the potential value of source-validity audits and the risks of carryover, omissions, automation bias, and use before prospective clinical validation.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release of data or models that have a high risk for misuse?

    Answer: [N/A]

    Justification: The release contains protocol and evaluation code. The underlying clinical datasets retain their credentialed access controls.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets used in the paper properly credited and are license and terms of use explicitly mentioned and respected?

    Answer: [Yes]

    Justification: External models, datasets, and evaluation resources are cited. Access follows the original PhysioNet and model-provider terms; figure-asset credits and licenses accompany the source files.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The protocol, controller, evaluation code, and supporting assets are documented in Sections 3–4, the appendices, and the public repository.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and details about compensation?

    Answer: [Yes]

    Justification: Appendix C gives the reader instructions, response definitions, blinding and judgment-locking sequence, and participation and compensation conditions.

15. **Institutional review board approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants and whether IRB approvals were obtained?

    Answer: [Yes]

    Justification: Appendix C describes institutional authorization and the recorded ethics/IRB determination for annotation. No new patients were enrolled; controlled MIMIC-CXR materials stayed in authorized storage.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods?

    Answer: [Yes]

    Justification: Sections 2–4 identify the fixed report generators and describe their role in the source-causal protocol and controller.

A Executable Controller Details

Feature map. Let $n(b)$ be candidate token count, $\tau(b)$ its detected temporal terms, $\rho_3(b,h)$ the precision of distinct candidate word trigrams against valid prior report h, and sim normalized lowercase sequence similarity. The shared feature vector is

$$\begin{aligned}\psi_i(b) = \big[&1,\ \log(1+n(b))/5,\ |P(b)|/10,\ |\tau(b)|/5,\ \rho_3(b,h_i^{\mathrm{text}}),\\ &\mathrm{sim}(b,h_i^{\mathrm{text}}),\ \mathrm{sim}(b,y_i^{\mathrm{cur}}),\ \mathrm{sim}(b,y_i^{\mathrm{real}}),\ \mathrm{sim}(b,y_i^{\mathrm{text}}),\\ &|P(b)\setminus P(y_i^{\mathrm{cur}})|/10,\ |P(y_i^{\mathrm{cur}})\setminus P(b)|/10,\\ &|P(b)\setminus P(h_i^{\mathrm{text}})|/10,\ |P(h_i^{\mathrm{text}})\setminus P(b)|/10,\ |P(b)\cap P(h_i^{\mathrm{text}})|/10\big]. \end{aligned} \tag{7}$$

Five one-hot indicators for $(\mathrm{cur}, b^{\cdot 2}, b^{\cdot 8}, b^{V0}, b^{V2})$ are appended in that order.

The implementation tokenizes report length with lowercase `[a-z]+` matches and computes sim with `difflib.SequenceMatcher` on the complete lowercase strings. Prior-copy precision is the fraction of distinct lowercase word trigrams in the candidate that also occur in the prior report. Temporal count uses the fixed lexicon {new, unchanged, stable, improved, improving, worse, worsened, worsening, increased, decreased, resolved, persistent, again}. $P(\cdot)$ is the rule-based 11-category extractor for atelectasis, cardiomegaly, consolidation, edema, fracture, mass/nodule, opacity, pleural effusion, pneumonia, pneumothorax, and support devices, with explicit local negation handling.

Proposal construction. Each of $b^{\cdot 2}$ and $b^{\cdot 8}$ is produced by a separately fitted averaged structured-perceptron reranker over exactly $(y^{\mathrm{cur}}, y^{\mathrm{real}}, y^{\mathrm{text}})$. Its feature map is the 14 shared text features above plus three base-arm indicators; its training target is the first base report maximizing $\mathrm{F1} - \lambda\,\mathrm{proxyFP} - .1\,\mathrm{proxyFN}$, with $\lambda = .2$ or $.8$. At inference the frozen proposal reranker receives no reference and returns the first score-maximizing base report unchanged, using the order current-only, real-history, prior-report-only.

The two verifier proposals use budget-specific logistic finding verifiers fitted on reference labels but evaluated at inference from positive-finding presence in those same three source-valid base reports and their fixed interactions; wrong-history arms are excluded. Each budget's verifier predicts an allowed finding set, takes current-only as its natural-language base, removes every base sentence containing a positive outside that set, and appends compact canonical statements for allowed findings still missing. Sentence appending uses a zero-token budget for b^{V0} and a two-token budget for b^{V2} relative to current-only. If filtering leaves no text, the implementation returns its canonical finding realization; this fallback can exceed the nominal append budget. Thus reference labels fit the frozen proposal models but are absent from every deployment-time proposal and final-router feature.

Fitting and routing pseudocode. For each evaluation panel E, remove every training study whose subject occurs in E. Initialize $w = 0$ and iterate 50 epochs over the remaining candidate sets. For case i, take the first candidate in the fixed order maximizing Equation (5); predict the first candidate maximizing $w^\top \phi(i,b)$; on a mistake update $w \leftarrow w + .05[\phi(i,b_i^*) - \phi(i,\widehat{b}_i)]$; average w over all steps. At deployment, generate the current-only, real-history, and prior-report-only reports; construct the four frozen proposals; compute the 19 features for every nonempty candidate; and return the first score maximizer unchanged.

B Aligned Direct Comparison with Current-Only

Table 10 reports paired 95% recipient-subject cluster-bootstrap intervals on the aligned 300-study subsets (5,000 resamples). Differences are computed before rounding.

Current-only/SCFR finding F1 is $.4193/.4220$, $.4139/.4224$, and $.4201/.4103$ for A–C; status accuracy is $.7018/.7052$, $.7039/.7045$, and $.7061/.7036$.

Table 8: Averaged coefficient vectors for Panels A–C in executable feature order. Values are rounded to six decimals; the accompanying machine-readable file provides full precision.

Feature	A	B	C
bias	.000000	.000000	.000000
$\log(1+\text{tokens})/5$	.119077	.020858	-.010136
positive findings $/10$	-.032463	.005488	-.046346
temporal terms $/5$	-.067226	.117758	.030072
prior trigram-copy precision	-.071506	-.119635	-.070541
similarity to valid prior	.180600	.322668	.032470
similarity to current-only	-.143765	-.094959	-.129449
similarity to real-history	-.003020	-.014375	.027465
similarity to prior-report-only	-.022558	-.007351	.007607
positive added vs. current $/10$	.060505	.167269	-.016453
positive removed vs. current $/10$	.092968	.161781	.029892
positive added vs. prior $/10$	-.112001	-.020995	-.172085
positive removed vs. prior $/10$	-.079538	-.026483	-.125739
positive overlap with prior $/10$	.079538	.026483	.125739
arm=current-only	.105806	.094293	.090477
arm=$b^{.2}$	-.057782	-.031646	-.047814
arm=$b^{.8}$	-.093879	-.051774	-.086300
arm=b^{V0}	.040699	.037749	.039914
arm=b^{V2}	.005155	-.048622	.003723

Table 9: Final-ranker fitting pools and route counts. Every evaluation has 300 studies; final-ranker fitting excludes evaluation subjects.

Panel	Fit rows	Eval. subjects	Cur	.2	.8	$V0$	$V2$
A	1,466	230	272	23	0	5	0
B	1,343	231	249	37	0	14	0
C	1,330	226	240	44	1	15	0

C Blinded Image-Grounded Reader Studies

C.1 Proxy calibration

An outcome-independent deterministic hash selects one study per subject from 18,716 eligible studies, leaving 5,750 subjects. Excluding 12 smoke subjects and a separately reserved 600-case probability cohort leaves the 5,138-subject challenge frame. Mutually exclusive strata are assigned in the following priority order, using the 11-category extractor:

1. Either wrong-history arm adds a positive category absent from current-only ($N = 3,643$).
2. Otherwise, current-only misses a positive current-reference category ($N = 1,160$).
3. Otherwise, random wrong-history changes the full finding-status vector relative to current-only ($N = 229$).
4. Otherwise, matched wrong-history changes that status vector ($N = 65$).
5. Otherwise, either wrong-history report differs from current-only after Unicode NFC and whitespace normalization, retaining case and punctuation ($N = 41$).

Twelve cases per stratum were deterministically hash-selected with seed 20260724. Inclusion probabilities are $12/N_s$ for stratum s. The prespecified sample therefore contained 60 unique-subject examinations and 660 fixed finding decisions. The reader marked 554 absent, 89 present, and 17 uncertain, with none missing or unassessable. Primary agreement excludes the 17 uncertain decisions, leaving 643 resolved labels. The reference-derived positive proxy has PPV .411, sensitivity .337, and specificity .922 against the reader's current-image judgments. Percentile intervals use 10,000 bootstrap replicates, resampling the 12 subject-unique cases with replacement separately within each stratum (seed 20260728). Inverse-probability/Hájek summaries use weights $N_s/12$ and the same

Table 10: SCFR-minus-current-only effects; brackets are paired 95% CIs.

Panel	Metric	SCFR−Current	95% CI
A	Finding F1	$+.0027$	$[-.0109, .0177]$
	Status accuracy	$+.0033$	$[-.0039, .0111]$
	FP/report	$-.0267$	$[-.0875, .0337]$
	FN/report	$.0000$	$[-.0525, .0489]$
B	Finding F1	$+.0085$	$[-.0083, .0255]$
	Status accuracy	$+.0006$	$[-.0072, .0082]$
	FP/report	$+.0433$	$[-.0194, .1093]$
	FN/report	$-.0233$	$[-.0741, .0298]$
C	Finding F1	$-.0097$	$[-.0288, .0091]$
	Status accuracy	$-.0024$	$[-.0118, .0066]$
	FP/report	$+.0233$	$[-.0572, .1096]$
	FN/report	$+.0100$	$[-.0511, .0695]$

Table 11: Primary SCFR-minus-matched-WH effects on the same controller subsets. Paired 95% subject-cluster intervals use 5,000 resamples.

Panel	Proxy FP/report	Proxy FN/report	Holm p
A	$-.3433$ $[-.454, -.237]$	$+.1700$ $[.075, .261]$	$\leq .0012$
B	$-.5500$ $[-.666, -.438]$	$+.1333$ $[.041, .229]$	$\leq .0012$
C	$-.4567$ $[-.576, -.342]$	$+.1500$ $[.059, .242]$	$\leq .0012$

stratified resampling with seed 20260729. Uncertain and unassessable labels are reported separately and excluded from resolved binary denominators.

Table 12: Image adjudication of positive additions versus current-only. Rates use resolved additions; brackets are 95% CIs.

Arm	Unsupported	Supported	Uncertain	Unsupported/resolved
Real history	7	11	0	38.9% $[14.3, 64.3]$
Matched wrong history	8	3	3	72.7% $[40.0, 100.0]$
Shuffled wrong history	8	5	2	61.5% $[41.6, 81.8]$

Inverse-probability weighting maps the challenge strata to the frozen 5,138-case challenge frame, not to all 18,716 studies. Matched wrong history exceeds real history by 25.9 unsupported-addition percentage points (95% CI 1.3–52.8), and exceeds current-only by $.404$ unsupported additions/case (95% CI .112–.710). Its combined-error contrast against current-only is .376/case (95% CI .051–.742).

C.2 Active-route comparison

Thirty subject-unique cases in which SCFR selected text different from current-only were sampled, 10 per panel. Anonymous A/B ordering was randomized and blinded. True history was clinically informative in 25/30. Preference was 13 for SCFR and 17 for current-only, with no ties (exact two-sided sign test $p = .585$). Active-route item intervals use 100,000 paired case-level bootstrap resamples (seed 20260728).

At case level, current-only/SCFR counts were 20/24 for any supported loss, 11/12 for any unsupported addition, and 19/17 for any important omission; McNemar p-values were $.454$, 1.000, and $.832$. SCFR expressed four correct and one partially correct temporal relation, with none incorrect; current-only had no assessable temporal relation in these cases.

C.3 Reader instructions and participation

The current-image task instructed the reader to review all available current-study views and select exactly one of *present*, *absent*, *uncertain*, or *unassessable* for each of 11 categories: atelectasis,

Table 13: Item-level active-route adjudication. Deltas are SCFR minus current-only; brackets are paired 95% CIs.

Endpoint	Current-only	SCFR	Mean delta [95% CI]
Supported findings represented	151	156	$+.167$ [$-.400, .733$]
Supported findings lost	47	50	$+.100$ [$-.733, .933$]
Unsupported additions	26	17	$-.300$ [$-.967, .333$]
Important omissions	20	17	$-.100$ [$-.433, .233$]

cardiomegaly, consolidation, edema, fracture, mass/nodule, opacity, pleural effusion, pneumonia, pneumothorax, and support device. Other clinically important findings were listed separately. Present meant supported by the current images; absent meant adequately assessable and not present; uncertain meant possible without a confident decision; and unassessable meant that image quality or available views prevented assessment. Blank fields did not mean absent. Prior images and text, candidate reports, references, system identities, proxy labels, and computational outputs were hidden. Responses were locked before reports or proxy labels were disclosed.

For the active-route task, the first form instructed the reader to open all current and verified true-prior images, without generated reports, and label history as *clinically informative*, *not clinically informative*, *uncertain*, or *unassessable*, with a brief clinical reason. This judgment was locked before the anonymous reports appeared. The second form presented randomly ordered reports A and B. The reader chose A, B, tie, or unassessable and explained the choice; listed represented supported findings, lost supported findings, unsupported additions, and important omissions separately for each report; and assigned temporal correctness as correct, partially correct, incorrect, not applicable, or unassessable. Each finding occupied one line, and an empty list meant no such findings. Instructions prioritized clinical content over writing style, specified that prior-image presence alone did not establish current-image presence, and required case order, report text, and form headers to remain unchanged.

The reader was a qualified radiologist; qualification and experience were documented separately from responses. The institution authorized the annotation and recorded an ethics/IRB determination before annotation. The reader contributed voluntarily without separate remuneration. Participation conditions and conflicts were documented before annotation. Controlled MIMIC-CXR data remained in authorized access-controlled storage.

D Simple Source-Control Baselines

These fixed controls use the complete 18,716-study snapshot, separately from the 300-study controller cohorts.

Table 14: Fixed controls on the complete 18,716-study snapshot. The oracle uses outcomes unavailable at deployment.

Method	Finding F1	Status	FP/report	FN/report
Always current-only	.4338	.7055	.7416	1.2122
No positive additions	.4267	.7073	.6146	1.2922
Block prior-only additions	.4109	.6963	.7931	1.2541
Outcome-conditioned oracle	.5345	.7361	.6750	.9603

Relative to current-only, no-positive-additions changes F1 by $-.0071$ (95% CI $[-.0093, -.0050]$), FP by $-.1270$ ($[-.1337, -.1202]$), and FN by $+.0800$ ($[.0748, .0852]$). Uniform deletion therefore exchanges fewer reference-discordant positives for more missed reference findings.

E Cross-Generator Out-of-Fold Evaluation

We used five subject-disjoint folds. For outer test fold f, fold $(f + 1) \bmod 5$ supplied a separate meta-gate set and the other three folds fitted the target-generator proposal models. The final 19-feature ranker and its Panel C coefficient vector remained frozen; the final ranker was not refitted to the target

generator. On the meta-gate fold, all four SCFR-minus-current point deltas had to meet the fixed bounds $(.02, .02, .10, .10)$ for F1, status, FP, and FN. Passing deployed the router on the outer test fold; failing returned current-only for that entire fold. Wrong-history outcomes did not determine this gate. This combines frozen final-router transfer with target-generator proposal fitting. Both generator panels remain within MIMIC-CXR.

Table 15: Cross-generator OOF results. Deltas are SCFR minus current-only; gate is deployed/fallback folds and routes are non-current selections.

Generator	Studies/subjects	Gate	Routes	ΔF1	ΔStatus	ΔFP	ΔFN
RadInfer	400/374	5/0	288	$+.1060$	$+.0382$	$-.2925$	$-.1200$
Libra	306/288	2/3	24	$-.0184$	$+.0009$	$-.0719$	$+.0752$

For RadInfer, current-only/SCFR values are $.1450/.2510$ F1, $.6073/.6455$ status, $.8125/.5200$ FP, and $1.7975/1.6775$ FN. Paired 95% CIs for the four deltas are $[.0634, .1460]$, $[.0271, .0490]$, $[-.3753, -.2101]$, and $[-.1897, -.0524]$. Its SCFR-minus-WH FP effect is $-.1975$ (95% CI $[-.3151, -.0851]$, $p = .0004$). For Libra, CIs are $[-.0336, -.0035]$ for F1 and $[.0298, .1234]$ for FN. RadInfer shows active routing; Libra shows the fallback/tradeoff boundary.

F Controller Hyperparameter Sensitivity

The final utility pair $(\lambda_{\mathrm{FP}}, \lambda_{\mathrm{FN}}) = (.2, .4)$ was selected on three development panels (the 400-study pilot and the parent cohorts for Panels B and C), then frozen before Panel A, which was generated using the recorded development-study exclusion list. Panels B and C provide development-reused descriptive evaluations, and Panel A provides the post-selection check. Final-ranker fitting excludes evaluation subjects within each panel.

The development utility sweep appears in Table 7.

Table 16: Component operating points: descriptive unweighted means over the three controller cohorts. Bold and italics mark the numerical best and second-best entries per metric.

Component	Finding F1 ↑	Status ↑	FP ↓	FN ↓
Current-only	.4177	.7039	.7244	1.2411
Recall-biased proposal	**.4242**	.7052	.7689	**1.2100**
Precision-biased proposal	.4060	.7059	.6178	1.3233
Strict verifier	.4127	**.7082**	**.6033**	1.2989
Soft verifier	.4113	*.7070*	*.6156*	1.3022
SCFR	*.4182*	.7044	.7378	*1.2367*

The proposal FN coefficient was .1. From FP weights $.2/.4/.6/.8$, the .2 and .8 endpoints were retained to provide recall- and precision-oriented candidates. From verifier append caps 0/2/4/8, caps 0 and 2 were retained: cap 4 increased FP on all three development panels and cap 8 increased FP wherever evaluated.

Table 17: Epoch sensitivity as descriptive unweighted means over the controller panels; the panels are not treated as one independent pooled cohort.

Epochs	F1	Status	FP	FN	Non-current routes
10	.419386	.705253	.738889	1.228889	142
25	.419290	.705152	.736667	1.230000	140
50	.418249	.704444	.737778	1.236667	139
100	.419099	.705152	.732222	1.234444	138

Halving the reporting bounds from $.02/.02/.10/.10$ to $.01/.01/.05/.05$ preserves the decision on all panels; worst direct changes are $-.0097$ F1, $-.0024$ status, $+.0433$ FP, and $+.0100$ FN. On held-out Panel A, the frozen choice gives $+.0027$ F1, $+.0033$ status, $-.0267$ FP, and $.0000$ FN versus current-only.

G Additional Source-Use Decomposition

Matched wrong history produces 18,451 positive additions versus current-only. The same finding appears in its supplied prior report for 40.8% (95% CI 40.1–41.6%). Exact sentence matches account for $.5\%$; near matches, including exact matches, for 2.2% (95% CI 1.9–2.4%). Same-prior-finding additions below the near-copy threshold account for 38.8%. Among 11,969 reference-discordant additions, 31.8% (95% CI 30.9–32.7%) contain the same prior finding and 2.0% (95% CI 1.8–2.3%) are near copies. These overlap statistics measure shared wording and finding categories; current-image support is assessed separately in the reader studies.

H Source-Control Construction

The donor builder requires subject identifiers and eligible prior context (at least one prior-image path or nonempty prior-report text). It excludes the target row and all rows from the target subject. The matched arm uses the target's normalized split. Among eligible donors it lexicographically minimizes (1) the symmetric-difference count between extracted positive findings in the target and donor *current reference reports*, (2) the absolute difference in prior-report word counts, and (3) a seeded random tie-breaker, followed by donor index. Matching thus uses a donor's current reference signature, rather than its historical report's finding signature. Prior images and prior text are transferred together, and donors may be reused. References are used for this offline reference-conditioned stress construction and for outcome measurement; the deployed router never receives them. The release includes the original donor-construction code and documents the manifest requirements and split handling.