LOCI: MITIGATING HALLUCINATIONS IN RADIOLOGY REPORTS WITH PHRASE-TRIGGERED COUNTERFACTUAL DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucinations remain a critical reliability issue for medical large vision-language models, particularly when generated text asserts clinical findings unsupported by the image. In chest X-ray report generation, mitigating such errors is challenging because clinically meaningful findings are often expressed as multi-token phrases whose interpretation depends on assertion scope, such as positive, negated, or uncertain. Existing training-free decoding methods typically operate at the token level or rely on global visual perturbations, making it difficult to selectively intervene on a specific clinical assertion. We introduce LOCI, a training-free, scope-aware decoding method that activates only when a candidate completes a clinical finding phrase in positive scope. For the matched finding, LOCI constructs a text-derived direction at the intervention layer, subtracts its projection from image-token hidden states in a temporary counterfactual branch, and reranks only plausible matched candidates according to the induced relative change in log probability. The edited branch is discarded after the current decoding step. Across three frozen 7B LVLMs and four chest X-ray datasets, LOCI obtains the best RadGraph F1 point estimate in all 12 backbone–dataset settings and the best CheXbert 5-category F1 result in 11 of 12 settings, although statistically resolved gains are concentrated in a subset of comparisons. On PadChest-GR, RadFact evaluation further shows higher logical F1 and fewer false positive findings across all three backbones. Mechanistic controls further show that the counterfactual signal is strongest for matched same-layer directions under the true-image condition. These results support phrase-triggered counterfactual sensitivity as a targeted inference-time signal for mitigating positive-finding hallucinations in frozen LVLMs. Code is available at `https://anonymous.4open.science/r/LOCI-837B/`.

1 INTRODUCTION

Large vision-language models (LVLMs) can generate fluent and semantically plausible text while still asserting content that is unsupported by the visual input, making hallucination a central reliability problem in multimodal generation (Liu et al., 2023; 2024b; Li et al., 2023a). This issue is particularly consequential in medical vision-language systems, where a single unsupported positive finding can alter the clinical interpretation of a generated report (Rohrbach et al., 2018; Liu et al., 2024a; Miura et al., 2021). We study this problem in X-ray report generation, where hallucinations frequently take the form of positive clinical findings that are not supported by the radiographic evidence. In this work, we refer to such errors as *positive-finding hallucinations*. Because direct visual-support annotations are unavailable for most report-generation benchmarks, our large-scale report-level evaluation necessarily relies on reference-based clinical agreement. We therefore distinguish reference-based factual agreement from direct image grounding throughout the paper.

Radiology makes this problem structurally different from generic object hallucination. The clinically meaningful unit is often not an isolated token but a finding phrase together with its assertion scope. For example, the token "effusion" becomes clinically consequential when it completes "pleural effusion", yet the same phrase should be treated differently in "no pleural effusion," or "cannot exclude pleural effusion." A decoder that penalizes individual finding tokens therefore risks suppressing valid negated or uncertain statements, while a global intervention may alter unrelated parts of the report. Effective

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.

mitigation should instead act selectively at the moment when a clinically meaningful positive assertion is about to be completed.

Training-free decoding is attractive in this setting because it can recalibrate frozen LVLMs without additional supervision or parameter updates (Li et al., 2023b; Leng et al., 2024; Huang et al., 2024). Existing approaches, however, generally rely on global or generic contrasts rather than the completion of a clinically meaningful assertion. The clinical structure above instead calls for an intervention that is phrase-level, assertion-aware, and local to the current decoding decision. Rather than globally suppressing a clinical concept, we ask whether a specific positive finding candidate depends on the corresponding image-side representation.

We therefore propose LOCI (LOcal Counterfactual Intervention), a scope-aware, phrase-triggered decoding method that uses concept removal as a local counterfactual probe. When a candidate is about to complete a clinical finding in positive scope, LOCI temporarily subtracts the projection along the matched phrase direction from image-token hidden states and measures the induced change in candidate probability. A large probability decrease indicates stronger dependence on the corresponding image-side representation, whereas little change suggests greater reliance on textual context or language prior. LOCI uses the relative change only to rerank plausible matched candidates. The counterfactual branch is immediately discarded, and decoding is unchanged at other steps.

We evaluate LOCI on chest X-ray report generation across multiple frozen LVLM backbones and datasets. LOCI consistently improves RadGraph F1 over standard decoding and yields substantial CheXbert gains in several settings, especially on Qwen2.5-VL. Although the improvements are not uniform across all metrics and come with a latency trade-off, the results suggest that phrase-triggered latent interventions are a practical and interpretable direction for reducing unsupported positive findings in medical LVLM generation.

Contributions. We make three contributions:

- We formulate radiology hallucination mitigation as a phrase-level, assertion-dependent decoding problem in which the relevant intervention event is the completion of a positive clinical finding rather than the generation of an isolated token.
- We introduce a local counterfactual decoding formulation in which matched concept removal is used only as a *temporary probe*. The edited representation is discarded after the current decoding step, and only the relative change in candidate probability is used to rerank plausible positive-scope finding completions.
- We provide broad evaluation across three frozen LVLMs and four chest X-ray datasets, combining reference-based clinical agreement, logical-entailment evaluation, and mechanistic controls. The results show improved clinical agreement and fewer reference-non-entailed positive findings without reducing the method to generic positive-finding suppression.

2 Related work

Radiology report generation and clinical hallucination. Radiology report generation has long faced a gap between fluent surface realization and clinically faithful content. Prior work improves factuality through task-specific architectures, factual rewards, and RadGraph-based semantic rewards (Miura et al., 2021; Delbrouck et al., 2022), while recent systems increasingly build on pretrained LLM and LVLM backbones such as R2GenGPT and LLaVA-Med (Wang et al., 2023; Li et al., 2023a). These systems are commonly evaluated with CheXbert and RadGraph in addition to lexical metrics (Smit et al., 2020; Jain et al., 2021). Unlike generic object hallucination benchmarks such as CHAIR, POPE, and HallusionBench (Rohrbach et al., 2018; Li et al., 2023c; Guan et al., 2024), radiology errors often involve multi-token findings whose meaning depends on assertion status. Radiology NLP has therefore long treated polarity as a first-class problem: systems such as NegBio (Peng et al., 2018) distinguish positive, negated, and uncertain findings. LOCI brings this assertion awareness into generation-time intervention, using online polarity to determine when a frozen decoder should be modified rather than applying assertion analysis only after generation.

Training-free decoding for hallucination mitigation. Inference-time methods improve grounding without updating model weights. VCD (Leng et al., 2024) contrasts predictions under original and perturbed visual inputs, while OPERA(Huang et al., 2024) uses attention-based over-trust penalties.

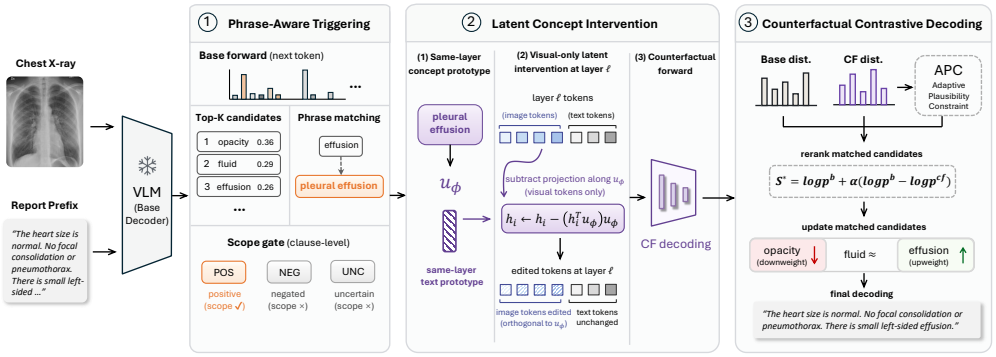


Figure 1: Overview of LOCI. A top-K phrase trigger and left-context scope gate identify positive finding completions. LOCI then subtracts a scaled projection along the matched phrase direction from image-token states at a fixed post-block residual hook and reranks only the matched candidates using counterfactual log-probabilities.

Recent methods also intervene directly in multimodal representations, including ProjectAway(Jiang et al., 2025a) and VTI (Liu et al., 2025); Med-VCD (Mahdavi et al., 2026) adapts visual contrastive decoding to medical LVLMs. These methods demonstrate the effectiveness of inference-time calibration, but generally rely on global visual perturbations, generic latent interventions, or step-level distributional contrasts. LOCI instead activates only when a candidate completes a specific clinical finding in positive scope and constructs a concept-matched counterfactual for that finding.

Representation intervention and concept erasure. LOCI is also related to representation-level concept removal, including INLP (Ravfogel et al., 2020) and LEACE (Belrose et al., 2023), which remove targeted information from hidden representations. We do not treat the projection operator itself as the main novelty. The key distinction is how the edited representation is used: LOCI discards the edited branch after the current decoding step and uses only the relative probability change as a local counterfactual signal for matched candidates, rather than decoding from the erased representation.

3 METHOD

Figure 1 summarizes LOCI. Given image I and prefix $Y_{<t}$, standard decoding produces the next-token distribution p_t^b. When a top-K candidate completes a positive-scope finding, LOCI constructs a same-layer text prototype, removes its projection only from image-token states, and uses the induced log-probability change to rerank plausible matched candidates. Otherwise decoding is unchanged, and the counterfactual branch is discarded after the current step. Let f_θ be a frozen autoregressive LVLM with vocabulary $\mathcal{V}$. Given image I and prefix $Y_{<t}$ at decoding step t, the base next-token distribution is $p_t^b(y) = P_\theta(y \mid I, Y_{<t}), y \in \mathcal{V}$. We describe greedy decoding for clarity. The same reranking scores can be used with other decoding policies. Repetition penalties, token filters, and stopping rules are applied identically to standard decoding and LOCI.

3.1 ONLINE PHRASE TRIGGER AND SCOPE GATE

Phrase trigger. LOCI inspects only a small candidate set from the base distribution, $\mathcal{C}_t = \mathrm{TopK}(p_t^b, K_{\mathrm{cand}})$. For each candidate token $c \in \mathcal{C}_t$, we detokenize the current prefix together with c, normalize the resulting text, and extract the current clause after the most recent sentence or adversative boundary. All matching is performed on detokenized text rather than raw subword IDs, so partial BPE pieces do not trigger until they complete a full phrase. Let $\mathcal{F}$ be a lexicon of canonical clinical findings, and let $\mathcal{A}(\phi)$ denote the alias set for finding $\phi \in \mathcal{F}$. For each candidate c, LOCI checks whether the current clause after appending c ends, on a word boundary, with an alias in $\mathcal{A}(\phi)$. When several alias strings match the same clause suffix, LOCI chooses the longest matched span before mapping it to the corresponding canonical finding. Location and laterality qualifiers are stripped when mapping surface spans to canonical findings, with exact rules provided in the appendix.

Scope gate. A phrase completion is clinically risky only when it is asserted positively. Therefore for each matched candidate, LOCI assigns an online polarity label from the left context in the current clause, $\sigma_t(c) \in \{\mathrm{POS}, \mathrm{NEG}, \mathrm{UNC}\}$. Only positive matches are intervened on; negated or uncertain matches are left to the base decoder. The gate uses precedence-ordered cue matching, $\mathrm{POS\text{-}override} > \mathrm{UNC} > \mathrm{NEG} > \mathrm{default\ POS}$, where sentence boundaries and adversatives reset scope, while commas and coordination remain in scope. Thus coordinated negations such as *no focal consolidation, edema, or pneumothorax* are skipped, whereas persistence statements such as *no change in pleural effusion* remain positive. Positive matched candidates are grouped by canonical finding:

$$\mathcal{M}_t(\phi) = \{c \in \mathcal{C}_t : \mu_t(c) = \phi,\ \sigma_t(c) = \mathrm{POS}\}, \qquad \mathcal{G}_t = \{(\phi, \mathcal{M}_t(\phi)) : \mathcal{M}_t(\phi) \neq \varnothing\}.$$

Groups are sorted by their best base score and truncated to at most $G_{\max}$ groups per step. If $\mathcal{G}_t$ is empty, LOCI exactly reduces to standard decoding.

3.2 Same-layer finding prototypes

For each canonical finding ϕ, LOCI constructs a concept direction in the same hidden space where the intervention will be applied. Let ℓ be the intervention layer, and let $g_\ell(\cdot)$ denote the post-block residual stream returned by decoder block ℓ, before the input normalization of block $\ell + 1$. This hook location is determined per backbone on validation data. To build the prototype for ϕ, we run a text-only forward pass using the same backbone-specific report-generation chat template and instruction prefix as test-time decoding, but without image tokens. The assistant continuation is set to the canonical phrase text of ϕ. Let $H^{(\ell)}_{\mathrm{text}}(\phi) \in \mathbb{R}^{n_\phi \times d}$ be the hidden states of the phrase tokens at the same hook location. The normalized finding direction is $\mathbf{u}_\phi = \mathrm{Pool}\Big(H^{(\ell)}_{\mathrm{text}}(\phi)\Big)$, $\hat{\mathbf{u}}_\phi = \frac{\mathbf{u}_\phi}{\|\mathbf{u}_\phi\|_2}$. We use final-token pooling by default and cache all prototypes before decoding. Because the report-generation prompt is fixed within each backbone, the same prototype bank is reused across report-generation datasets for that backbone.

3.3 Visual-only latent intervention

For each triggered finding group $(\phi, \mathcal{M}_t(\phi))$, LOCI performs one counterfactual forward pass with a visual-only hook at layer ℓ. Let $H^{(\ell)}_t = [\mathbf{h}^{(\ell)}_{t,1}, \ldots, \mathbf{h}^{(\ell)}_{t,L}]^\top \in \mathbb{R}^{L \times d}$ be the layer-ℓ hidden states for the current multimodal prefix. We define a binary mask $m^{\mathrm{vis}}_i \in \{0, 1\}$, where $m^{\mathrm{vis}}_i = 1$ iff position i corresponds to an image token from the original prompt. All non-image positions are left unchanged. For the matched finding ϕ, LOCI subtracts the projection of each image-token hidden state onto the phrase direction:

$$\tilde{\mathbf{h}}^{(\ell,\phi)}_{t,i} = \mathbf{h}^{(\ell)}_{t,i} - \lambda\, m^{\mathrm{vis}}_i \big(\mathbf{h}^{(\ell)\top}_{t,i}\, \hat{\mathbf{u}}_\phi\big)\, \hat{\mathbf{u}}_\phi, \tag{1}$$

where $\lambda \geq 0$ controls the intervention strength. This operation removes only the component aligned with the matched clinical concept and only from image-token states. Model parameters, text-token hidden states, and the generated prefix are unchanged. The hooked counterfactual forward pass yields a next-token distribution $p^{\mathrm{cf},\phi}_t$ over $\mathcal{V}$. Intuitively, if the probability of a finding candidate decreases after subtracting the matched concept direction from image-token states, the candidate is sensitive to the corresponding image-token representation. If it remains probable under the counterfactual intervention, LOCI treats the candidate as more likely to be supported by language priors or contextual co-occurrence.

3.4 Local counterfactual reranking and decoding schedule

LOCI uses the counterfactual next-token distribution only for matched positive finding candidates. It does not modify the scores of the full vocabulary. First, LOCI applies an adaptive plausibility constraint: $a_t(y) = \mathbf{1}\big[p^b_t(y) \geq \alpha_{\mathrm{apc}} \max_{v \in \mathcal{V}} p^b_t(v)\big]$. This prevents the counterfactual branch from promoting candidates that are implausible under the base model. Let $\mathcal{U}_t = \bigcup_{(\phi, \mathcal{M}_t(\phi)) \in \mathcal{G}_t} \mathcal{M}_t(\phi)$ be the set of all processed matched candidates, and let ϕ_y denote the canonical finding associated with candidate $y \in \mathcal{U}_t$. For each matched candidate, LOCI computes the counterfactual log-probability difference $m_t(y, \phi) = \log p^b_t(y) - \log p^{\mathrm{cf},\phi}_t(y)$. Both probabilities are computed from the complete next-token distribution over $\mathcal{V}$. The default margin is therefore a normalized log-probability change;

Table 1: Chest X-ray report-generation results on four datasets. R-F1 denotes RadGraph F1, and 14c-F1/5c-F1 denote CheXbert positive-class macro-F1 (%). Bold and underlined values indicate the best and second-best results within each backbone, respectively.

		MIMIC-CXR			IU-Xray			CheXpert Plus			PadChest-GR		
Backbone	**Method**	**R-F1**	**14c-F1**	**5c-F1**	**R-F1**	**14c-F1**	**5c-F1**	**R-F1**	**14c-F1**	**5c-F1**	**R-F1**	**14c-F1**	**5c-F1**
LLaVA-Med 7B	Standard (Baseline)	7.98	12.10	16.80	14.65	7.32	4.85	7.93	4.78	4.42	1.32	5.98	5.64
	VCD	6.16	9.78	8.98	12.95	8.58	**7.11**	6.82	4.15	4.87	1.12	6.13	5.44
	Med-VCD	6.87	9.27	8.39	12.29	7.17	4.95	7.53	4.55	5.12	1.20	6.05	5.71
	ProjectAway	8.15	11.70	13.78	10.46	7.50	6.68	7.41	4.95	5.05	1.35	6.35	5.29
	VTI	7.36	9.20	8.18	9.57	7.39	5.12	7.88	4.47	6.03	1.25	6.20	4.88
	OPERA	4.48	7.26	11.31	15.34	6.03	4.00	5.10	4.15	5.23	1.06	5.85	4.35
	LOCI (Ours)	**9.41**	**12.38**	**18.73**	**16.33**	**8.66**	6.97	**8.46**	**5.06**	**8.26**	**1.39**	**6.39**	**5.88**
Hulu-Med 7B	Standard (Baseline)	20.13	20.63	29.96	27.79	7.11	5.93	16.35	17.11	19.96	**2.01**	15.04	12.43
	VCD	19.16	18.58	28.12	25.13	6.82	5.11	14.88	16.29	18.04	1.77	14.02	11.52
	Med-VCD	18.72	18.14	27.44	26.41	6.91	5.28	15.02	16.48	18.33	1.63	13.78	11.19
	ProjectAway	20.08	20.11	29.21	27.14	7.03	5.81	16.12	17.04	19.82	1.94	14.88	12.16
	VTI	19.53	17.92	26.42	25.82	6.46	4.93	14.67	16.08	17.91	1.58	13.51	10.97
	OPERA	18.49	19.26	27.88	26.93	6.79	5.46	15.79	16.81	19.03	1.81	14.16	11.74
	LOCI (Ours)	**20.57**	**21.09**	**30.64**	**30.53**	**7.83**	**9.11**	**16.53**	**18.17**	**22.44**	2.05	**15.74**	**13.05**
Qwen2.5-VL 7B	Standard (Baseline)	9.85	11.33	17.55	27.65	7.40	7.52	9.48	12.14	18.77	3.31	4.64	4.37
	VCD	10.35	10.80	18.10	25.85	7.02	7.77	7.65	13.10	19.20	3.12	5.15	5.11
	Med-VCD	9.33	11.67	20.88	18.24	7.45	7.67	9.20	11.95	19.50	3.20	5.51	6.23
	ProjectAway	8.40	11.90	21.50	22.35	6.78	5.42	8.85	13.25	18.50	3.35	5.62	6.85
	VTI	9.70	12.10	17.12	26.45	6.34	4.39	9.30	**14.05**	19.10	3.25	4.53	4.95
	OPERA	9.65	12.05	17.90	26.32	6.44	3.18	7.15	13.40	20.80	3.80	5.91	7.95
	LOCI (Ours)	**11.07**	**13.53**	**22.88**	**27.92**	**7.78**	**7.82**	**9.78**	13.80	**21.92**	**6.56**	**6.18**	**8.48**

Appendix Table 13 reports raw-logit and centered-logit variants. LOCI then defines the reranking score

$$S_t(y) = \begin{cases} \log p_t^b(y) + \alpha\, m_t(y, \phi_y), & y \in \mathcal{U}_t,\ a_t(y) = 1, \\ \log p_t^b(y), & \text{otherwise.} \end{cases} \tag{2}$$

A positive counterfactual difference indicates that removing the matched image-side concept decreases the candidate probability, and therefore increases its final score. A negative difference indicates that the candidate becomes relatively more probable after concept removal and is therefore downweighted. All unmatched candidates retain their base log-probabilities. After reranking, LOCI applies the same token filters, repetition handling, and stopping rules as standard decoding and emits $y_t^\star = \arg\max_{y \in \mathcal{V}} S_t(y)$. Since greedy decoding depends only on candidate ranking, the scores S_t do not need to be explicitly renormalized. The selected token is appended to the prefix, and the base branch continues with ordinary KV-cache decoding. The counterfactual branch is used only to score the current step and is not carried forward.

Cost. LOCI adds no trainable parameters. Extra computation occurs only at positive phrase-completion steps, with approximate cost $C_{\text{LOCI}} \approx C_{\text{base-cache}} + \rho\, \bar{G}\, C_{\text{cf-full}}$, where ρ is the trigger rate and $\bar{G} \le G_{\max}$ is the average number of processed finding groups.

4 Experiments

We evaluate LOCI on chest X-ray report generation with frozen LVLMs. The experiments ask whether LOCI improves reference-based clinical agreement, whether the gains are more than generic conservatism, and whether evaluation against radiologist-curated annotations shows fewer unsupported positive findings. MIMIC-CXR, IU-Xray, and CheXpert Plus provide reference-based RadGraph and CheXbert evaluation, whereas PadChest-GR provides radiologist-curated findings and location annotations. Ablations are used as mechanistic diagnostics rather than independent clinical adjudication.

4.1 Experimental setup

Datasets. We evaluate on four chest X-ray report-generation datasets: MIMIC-CXR (Johnson et al., 2019), IU-Xray (Demner-Fushman et al., 2016), CheXpert Plus (Chambon et al., 2024), and PadChest-GR (Castro et al., 2025). For PadChest-GR, we use the official English-translated version. Note that this dataset's unique reference format leads to lower reference-based metric scores across all methods (see Appendix A.1 for details). The test sets contain 5,159 MIMIC-CXR image–report

Table 2: Assertion behavior on MIMIC-CXR with Qwen2.5-VL. Len is the average report length in words.

Method	Trigger	Scope	Image-token edit	Concept-specific	Relative CF	R-F1	14c-F1	5c-F1	Pos-Prec	Pos-Rec	Len
Standard decoding	×	×	×	×	×	9.85	11.33	17.55	16.34	20.61	46.2
Tuned conservative prompt	×	×	×	×	×	10.24	11.95	17.94	17.83	18.11	43.8
Scope-aware lexicon penalty	✓	✓	×	×	×	10.12	11.85	17.71	18.62	16.91	44.5
Blank-image contrastive	✓	✓	×	×	✓	10.41	12.90	20.08	20.42	19.82	45.1
Random-direction intervention	✓	✓	✓	×	✓	9.82	11.28	17.45	16.21	20.45	45.8
Scope-aware direct erasure	✓	✓	✓	✓	×	10.88	12.77	17.98	17.48	18.79	44.9
LOCI	✓	✓	✓	✓	✓	**11.07**	**13.53**	**22.88**	**24.16**	**22.40**	45.5

pairs, 639 IU-Xray pairs, 202 CheXpert Plus pairs, and 915 PadChest-GR pairs. When both Findings and Impression sections are available, the reference report is their concatenation. For multi-view studies, each image is processed independently and evaluated against the corresponding study report. Appendix D.6 reports a single-view versus multi-view analysis.

Backbones and decoding methods. We evaluate three frozen 7B LVLMs: LLaVA-Med v1.5 (Li et al., 2023a), Hulu-Med 7B (Jiang et al., 2025b), and Qwen2.5-VL-7B-Instruct (Bai et al., 2025). All model weights remain frozen. We compare LOCI with standard greedy decoding and training-free decoding baselines: VCD (Leng et al., 2024), Med-VCD (Mahdavi et al., 2026), OPERA (Huang et al., 2024), ProjectAway (Jiang et al., 2025a), and VTI (Liu et al., 2025). All methods use the same prompt semantics, image preprocessing, maximum generation length, repetition penalty, token filters, and stopping rules.

Reference-based metrics. We report RadGraph F1 (Jain et al., 2021) and CheXbert-based clinical agreement (Smit et al., 2020). For CheXbert, POS is treated as the positive class, while NEG, UNC, and BLANK are merged as non-positive. We compute positive-class macro-F1 over all 14 CheXbert observations, denoted 14c-F1, and over the five-observation subset defined in Appendix B.4, denoted 5c-F1. Positive precision and recall use the same five-observation binarization. ROUGE-L (Lin, 2004) is reported only as a lexical-overlap diagnostic.

RadFact logical-entailment metrics on PadChest-GR. We use the official RadFact toolkit (Bannur et al., 2024) to evaluate textual entailment between generated reports and the radiologist-curated PadChest-GR references. RadFact logical precision measures the fraction of generated finding sentences that are entailed by the reference, while logical recall measures the fraction of reference finding sentences that are entailed by the generated report.

Selection and uncertainty. All prompts, lexicons, prototype banks, hyperparameters, and control settings are selected using validation data disjoint from the test sets and are frozen before test evaluation. For each backbone, one LOCI intervention layer and one hyperparameter configuration are selected on the MIMIC-CXR validation split and then reused unchanged on all four test datasets. Unless otherwise stated, confidence intervals use 10,000 paired bootstrap resamples, clustered by study when study identifiers are available. Layer and hyperparameter sweeps are reported in Appendix Figures 5 and 6.

4.2 Reference-based report-level agreement

Table 1 reports reference-based results across four datasets and three frozen LVLMs. LOCI obtains the highest RadGraph F1 in all 12 backbone–dataset conditions, the highest CheXbert 5c-F1 in 11 conditions, and second-best 5c-F1 in the remaining one. Relative to standard decoding, LOCI improves every reported RadGraph, 14c-F1, and 5c-F1 point estimate.

The largest gains occur for Qwen2.5-VL: on MIMIC-CXR, RadGraph F1 improves from 9.85 to 11.07 and 5c-F1 from 17.55 to 22.88; on PadChest-GR, RadGraph F1 improves from 3.31 to 6.56 and 5c-F1 from 4.37 to 8.48. Paired 95% confidence intervals for LOCI minus standard decoding exclude zero in 6 of 12 RadGraph comparisons and 6 of 12 5c-F1 comparisons. A conservative bootstrap comparison against the strongest competing training-free decoder is statistically resolved in 5 of 12 RadGraph and 4 of 12 5c-F1 conditions (Appendix Table 24). Thus, the main evidence is consistent first-place reference-based performance, with statistically resolved superiority concentrated in a subset of settings.

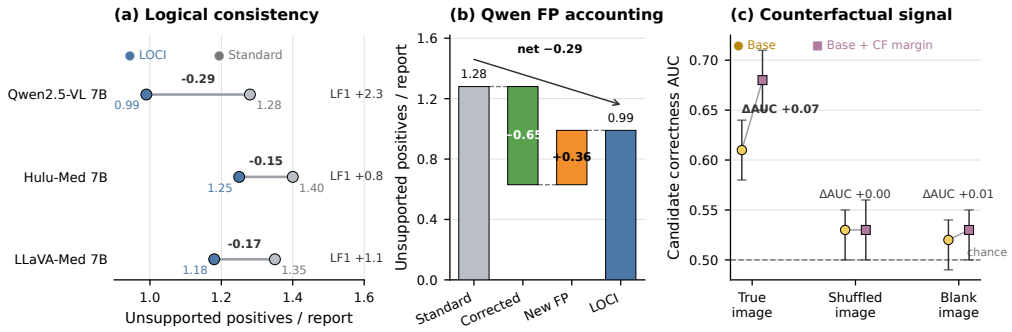


Figure 2: RadFact logical-entailment and counterfactual diagnostics. **(a)** Reference-non-entailed positive findings and logical F1 on PadChest-GR. **(b)** Non-entailed-positive transition accounting for Qwen2.5-VL. **(c)** Candidate-correctness AUC with and without the counterfactual margin under true, shuffled, and blank images. Error bars denote 95% study-clustered bootstrap confidence intervals.

Table 3: RadFact logical-entailment evaluation on PadChest-GR. Panel A compares standard decoding and LOCI across backbones; Panel B reports Qwen2.5-VL diagnostic controls. LPrec and LRec denote RadFact logical precision and logical recall; LF1 is their harmonic mean.

Backbone	Method	LPrec ↑	LRec ↑	LF1 ↑	UnsupPos/report ↓
A. Standard decoding versus LOCI across backbones					
LLaVA-Med	Standard decoding	17.5	19.0	18.2	1.35
	LOCI	**19.8**	**18.8**	**19.3**	**1.18**
Hulu-Med	Standard decoding	16.8	18.1	17.4	1.40
	LOCI	**18.5**	**17.9**	**18.2**	**1.25**
Qwen2.5-VL	Standard decoding	20.0	23.2	21.5	1.28
	LOCI	**24.5**	**23.2**	**23.8**	**0.99**
B. Qwen2.5-VL diagnostic controls (evaluated with RadFact)					
Qwen2.5-VL 7B	Tuned conservative prompt	22.1	20.5	21.3	1.10
	Scope-aware lexicon penalty	22.5	19.8	21.1	1.05
	Blank-image contrastive	23.0	21.1	22.0	1.08
	Scope-aware direct erasure	23.4	22.1	22.7	1.04
	Random-direction intervention	20.4	22.9	21.6	1.22

4.3 How LOCI Achieves Gains: Targeted Positive-Finding Reduction

Because LOCI intervenes on positive-scope finding completions, we analyze whether its gains stem from targeted adjustments rather than generic conservatism. While LOCI does reduce the total number of positive assertions, it does so in a way that improves both precision and recall, unlike simpler baseline methods.

On the MIMIC-CXR test set with Qwen2.5-VL, LOCI reduces the total number of predicted positive findings for the main 5 CheXbert categories by approximately 28% (Appendix Table 30). Similarly, on PadChest-GR, the total number of positive findings per report decreases by about 19% (from 1.59 to 1.30, Tables 3 and 9), almost entirely by reducing reference-non-entailed positives while preserving entailed ones. Crucially, as shown in Table 2, this reduction leads to a marked improvement in both positive precision (16.34 to 24.16) and positive recall (20.61 to 22.40), boosting 5c-F1 from 17.55 to 22.88. In contrast, a tuned conservative prompt or a scope-aware lexicon penalty primarily sacrifices recall for precision. This demonstrates that LOCI's mechanism is more sophisticated than generic positive suppression. It selectively prunes less reliable assertions while preserving or even recovering correct ones, as further detailed in the transition analysis in Appendix Table 6.

4.4 RadFact Logical-Entailment Evaluation on PadChest-GR

We next evaluate reference-based factual consistency on PadChest-GR using RadFact logical precision and recall (Bannur et al., 2024). RadFact tests entailment in both directions: logical precision measures whether generated finding sentences are supported by the radiologist reference, whereas logical recall measures whether reference finding sentences are recovered by the generated report. No spatial grounding information is used in these metrics.

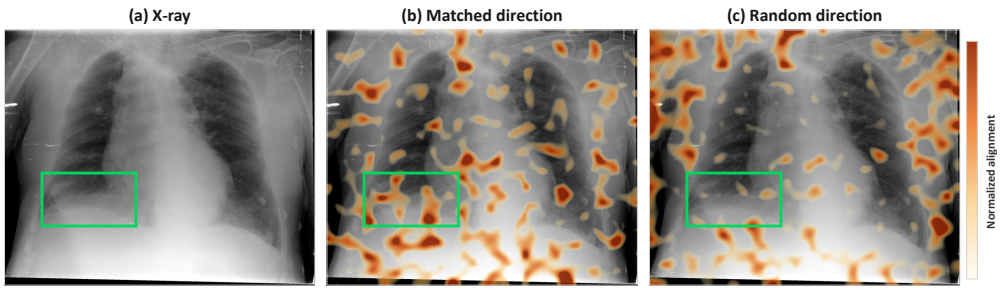


Figure 3: Representative spatial alignment for atelectasis: **(a)** radiologist annotation, **(b)** matched atelectasis direction, and **(c)** random-direction control. Warmer colors indicate higher alignment.

As shown in Table 3 (Panel A) and Fig. 2(a), LOCI improves logical entailment consistency over standard decoding. For Qwen2.5-VL, logical F1 increases from 21.5 to 23.8, primarily through an increase in logical precision from 20.0 to 24.5 while logical recall remains unchanged. Similar improvements in logical precision and F1 are observed for LLaVA-Med and Hulu-Med. LOCI also reduces the number of generated positive findings that are not entailed by the radiologist reference across all three backbones. For Qwen2.5-VL, Fig. 2(b) further decomposes this reduction: LOCI corrects 0.65 non-entailed positive findings per report while introducing 0.36 new ones, yielding a net reduction of 0.29 per report (Appendix Table 9).

Panel B further shows that the full relative-counterfactual formulation outperforms the diagnostic controls. In particular, scope-aware direct erasure improves over standard decoding but remains below LOCI, especially in logical precision. These results support the value of using the representational change as a reranking signal rather than directly decoding from the erased representation. Importantly, because this evaluation is based on textual entailment, these results should be interpreted as improved consistency with the radiologist reference rather than direct evidence that individual generated findings are visually grounded in the image.

Table 4: Internal LOCI ablations on MIMIC-CXR with Qwen2.5-VL. Each row changes one LOCI design choice while keeping the remaining configuration fixed. Simple clinical controls such as conservative prompting and lexicon penalties are reported separately in Table 2.

Variant	R-F1	14c-F1	5c-F1	R-L
Full LOCI	**11.07**	**13.53**	**22.88**	**35.58**
Ablating Reranking Strategy				
Rerank full vocabulary (instead of local)	10.71	12.88	21.30	35.31
Ablating Plausibility Constraint				
w/o Plausibility constraint	10.95	13.41	22.50	35.52
Fixed top-p constraint	11.00	13.44	22.58	35.54
Ablating Trigger Mechanism				
w/o Scope gate (positive-scope awareness)	10.55	12.15	19.80	35.40
Token-level trigger (instead of phrase-level)	10.38	12.42	20.35	35.32
Random-step trigger	9.92	11.48	17.95	35.18
Standard decoding (ablates all components)	9.85	11.33	17.55	35.20

4.5 Ablations and Mechanistic Evidence

We next test whether LOCI depends on its intended components: phrase-level triggering, assertion-aware scope gating, matched same-layer concept directions, and image-token counterfactual intervention. These analyses are mechanistic diagnostics rather than independent clinical adjudication.

Table 5: Concept direction ablation on MIMIC-CXR with Qwen2.5-VL. Full LOCI uses final-token same-layer prototypes with ℓ_2 normalization.

Direction	R-F1	14c-F1	5c-F1	R-L
Standard decoding	9.85	11.33	17.55	35.20
Full LOCI	**11.07**	**13.53**	**22.88**	**35.58**
mean pooling	10.91	13.31	22.28	35.54
unnormalized	10.58	12.84	20.96	35.41
LM-head direction	10.30	12.50	20.15	35.35
Input-embedding direction	10.12	12.05	19.34	35.27
Random direction	9.82	11.28	17.45	35.15
Wrong-concept direction	9.76	11.12	17.21	35.11

Phrase-level triggering, scope gating, and local reranking. Table 4 shows that removing scope gating, replacing phrase-completion triggers with token-level triggers, or triggering at random steps substantially reduces 5c-F1. Reranking the full vocabulary and weakening the plausibility constraint also underperform full LOCI. The largest drops occur when phrase-level or assertion-aware modeling is removed.

Matched directions and image-side intervention. Table 5 shows that normalized same-layer phrase prototypes outperform LM-head, input-embedding, random, and wrong-concept directions. Prototype probes, controlled blank/shuffled-image tests, intervention-scope ablations, and energy-matched random/wrong-concept controls further support that the useful signal comes from matched image-token states rather than arbitrary perturbations (Appendix Tables 14,15, 16, and 17).

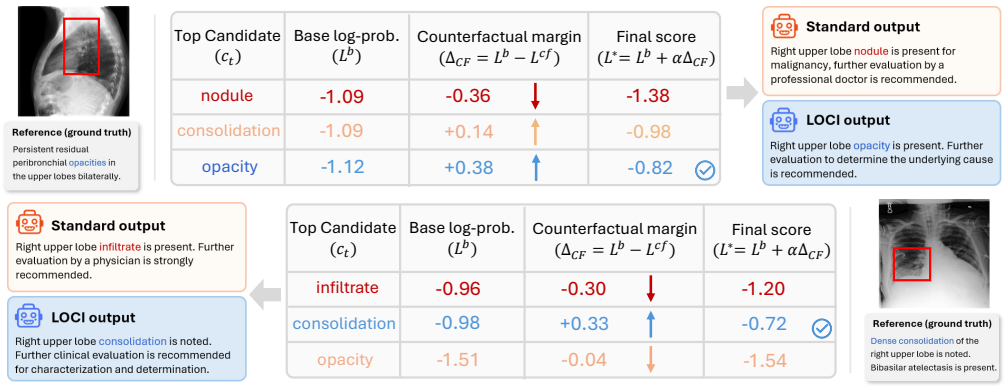


Figure 4: Qualitative examples of LOCI reranking. Red marks reference-inconsistent positive findings from standard decoding, and blue marks LOCI alternatives. The table rows show base log-probabilities, counterfactual margins, and final reranked scores. We use L to denote log-probabilities.

Coverage, spatial alignment, and score form. On PadChest-GR, replaying standard reports through the frozen matcher shows that 54.9–71.5% of unsupported positives are exposed to a positive-scope LOCI trigger (Appendix Table 11). Matched same-layer directions are also more spatially aligned with radiologist boxes than wrong-concept directions (Appendix Table 12). Figure 3 provides a representative example in which the matched atelectasis direction is more concentrated within the annotated region than the random-direction control. Finally, margin AUC, normalization variants, low-cost projection-only reranking controls, and a no-op $\lambda = 0$ sanity check are reported in Appendix Tables 13, 18, 19, and 20. Complementing these component ablations, Fig. 2(c) summarizes the image-dependent counterfactual-margin diagnostic. Adding the LOCI counterfactual margin improves candidate-correctness discrimination under the true-image condition, while providing little incremental value under shuffled or blank images.

4.6 Qualitative Examples and Efficiency

Qualitative examples. Figure 4 illustrates two phrase-completion events in which LOCI changes the local candidate ranking while leaving most of the surrounding report unchanged. In each example, the counterfactual margin alters the relative scores of plausible finding candidates and changes the token selected by standard decoding.

Efficiency. LOCI is invoked at only a small subset of decoding steps. On MIMIC-CXR, the three backbones average 2.18–3.05 positive triggers and 2.39–3.42 counterfactual forward passes per report. LOCI changes the standard greedy decision at 18.5–25.6% of positive-triggered steps (Appendix Table 7). The method is therefore sparse at the sequence level, although each processed finding group still requires a full counterfactual forward pass. In batch-1 measurements on an RTX 4090, LOCI requires 2.15–2.90× the latency of standard cached decoding, as detailed in our efficiency analysis in Appendix Table 8.

5 Discussion and Limitations

LOCI provides a training-free, phrase-triggered counterfactual decoding method for mitigating positive-finding hallucinations in chest X-ray reports. Its main limitation is computational cost: although interventions are sparse, each triggered finding requires a full counterfactual forward pass, resulting in 2.15–2.90× standard-decoding latency (Appendix Table 8). In addition, LOCI relies on simple text-derived concept directions, and the modest predictive value of the counterfactual margin (Appendix Table 18) suggests that its gains may partly reflect local regularization rather than precise visual grounding. LOCI may also replace one unsupported finding with another, a failure mode not fully captured by current automatic metrics. Nevertheless, its consistent improvements across multiple frozen LVLMs and datasets support phrase-triggered counterfactual sensitivity as a useful inference-time signal.

AI USE STATEMENT

We used generative AI tools solely to aid and polish the writing of this paper, including improvements to grammar, clarity, concision, and formatting. We did not use generative AI tools for research ideation, drafting scientific sections, generating synthetic data, or proving mathematical claims. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, 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/2406.04449.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

Daniel C. Castro, Aurelia Bustos, Shruthi Bannur, Stephanie L. Hyland, Kenza Bouzid, Maria Teodora Wetscherek, Maria Dolores Sánchez-Valverde, Lara Jaques-Pérez, Lourdes Pérez-Rodríguez, Kenji Takeda, José María Salinas, Javier Alvarez-Valle, Joaquín Galant Herrero, and Antonio Pertusa. PadChest-GR: A bilingual chest x-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7), 2025. doi: 10.1056/AIdbp2401120. arXiv:2411.05085.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats, 2024. URL https://arxiv.org/abs/2405.19538.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360, 2022.

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

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *ICLR*, 2025a. arXiv:2410.02762.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, Fudong Zhu, Kenli Li, Weidi Xie, Jimeng Sun, Jian Wu, and Zuozhu Liu. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding, 2025b. URL https://arxiv.org/abs/2510.08668.

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

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.

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. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023a.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 12286–12312, 2023b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023c.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024b.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 72402–72419, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b4008025c2182bfe16fcc8566ee14d64-Paper-Conference.pdf.

Zahra Mahdavi, Zahra Khodakaramimaghsoud, Hooman Khaloo, Sina Bakhshandeh Taleshani, Erfan Hashemi, Javad Mirzapour Kaleybar, and Omid Nejati Manzari. Med-vcd: Mitigating hallucination for medical large vision language models through visual contrastive decoding. *Computers in Biology and Medicine*, 200:111347, January 2026. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2025.111347. URL http://dx.doi.org/10.1016/j.compbiomed.2025.111347.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5288–5304, 2021. doi: 10.18653/v1/2021.naacl-main.416. URL https://aclanthology.org/2021.naacl-main.416/.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188, 2018.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7237–7256, 2020.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1500–1519, 2020.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023.

A Additional Report-Level Analyses

This appendix provides implementation and evaluation details that are omitted from the main paper for space. This section provides supplementary tables and figures supporting the main report-level results and efficiency analysis discussed in Sections 4.2, 4.3, and 4.6.

Table 6: Selected CheXbert assertion-state transitions on MIMIC-CXR with Qwen2.5-VL. Percentages are conditioned on the standard-decoding and reference states shown in the first column. The final three rows are average newly introduced reference-discordant assertions per report.

Standard/reference condition	LOCI outcome	Value
Base POS, reference BLANK	BLANK	43.1%
Base POS, reference BLANK	NEG	11.2%
Base POS, reference BLANK	UNC	7.2%
Base POS, reference BLANK	POS	38.5%
Base POS, reference POS	POS preserved	88.2%
Base POS, reference POS	Demoted to NONPOS	11.8%
Base NONPOS, reference POS	Recovered as POS	5.3%
New reference-discordant assertion	POS/report	0.18
New reference-discordant assertion	UNC/report	0.05
New reference-discordant assertion	NEG/report	0.07

Table 7: LOCI trigger statistics on the MIMIC-CXR test split. Counts are averaged per generated report. A match is a top-K candidate that completes a clinical finding phrase before scope filtering. POS matches trigger LOCI, while NEG and UNC matches are skipped by the scope gate. Changed decisions denote the percentage of POS-triggered steps where LOCI selects a different token from standard greedy decoding. CF fwd/report is the average number of counterfactual forward passes per report.

Backbone	Dataset	Tokens/report	Matches/report	Triggered steps/report	NEG skip/report	UNC skip/report	Changed dec. (%)	CF fwd/report
LLaVA-Med 7B	MIMIC-CXR	75.4	8.21	3.05	2.58	0.47	22.8	3.42
Hulu-Med 7B	MIMIC-CXR	63.8	6.55	2.18	2.12	0.42	18.5	2.39
Qwen2.5-VL 7B	MIMIC-CXR	71.6	9.08	2.84	3.18	0.54	25.6	3.16

Table 8: Inference efficiency on MIMIC-CXR. Latency is normalized to standard decoding within each backbone. All runs use batch size 1 on an RTX 4090.

Backbone	Method	Latency ↓	Peak VRAM (GB) ↓
LLaVA-Med 7B	Standard	1.00	15.40
	VCD	4.72	16.80
	Med-VCD	5.03	15.80
	OPERA	4.61	16.40
	LOCI	2.85	17.85
Hulu-Med 7B	Standard	1.00	15.20
	VCD	3.08	16.60
	Med-VCD	3.27	15.60
	OPERA	3.20	16.10
	LOCI	2.15	17.60
Qwen2.5-VL 7B	Standard	1.00	16.00
	VCD	3.78	17.40
	Med-VCD	4.05	16.40
	OPERA	3.96	17.00
	LOCI	2.90	18.20

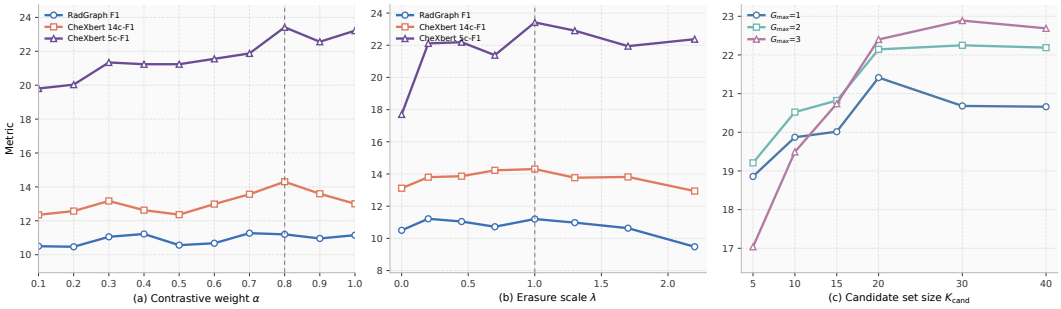


Figure 5: Hyperparameter sensitivity analysis on the Qwen2.5-VL / MIMIC-CXR validation split. Panels (a) and (b) vary one hyperparameter at a time while keeping others fixed. Panel (c) shows a two-factor sensitivity analysis for the candidate set size (K_{cand}) and the maximum group budget ($G_{\max}$).

A.1 Detailed RadFact Logical Evaluation on PadChest-GR

As noted in Section 4.1, we use the official English-translated version of the PadChest-GR dataset and its associated test split. The reference "reports" in this dataset are provided as a list of grounded sentences rather than a narrative paragraph. For reference-based metrics like RadGraph and CheXbert, we construct the reference_text by concatenating these grounded sentences. This format mismatch between the generated narrative reports and the structured sentence-based references explains the lower absolute scores observed for all methods on this dataset compared to MIMIC-CXR or IU-Xray. Our RadFact logical-entailment evaluation operates at the finding-sentence level and therefore provides a semantic entailment-based comparison between generated and radiologist-reference findings.

The following tables present detailed results and robustness checks for the positive-finding evaluation on PadChest-GR, as referenced in Section 4.4.

Table 9: RadFact-based error transition analysis on PadChest-GR with Qwen2.5-VL, derived from the RadFact results (Table 3). Percentages use transition-specific denominators: correction rates are normalized by the corresponding standard-decoding error set, preservation/suppression rates by standard-decoding grounded positives, new-error rates by LOCI errors, and recovery rates by grounded positives missed by standard decoding.

Transition type	Count/report	Percentage
Unsupported positive corrected	0.65	50.8
Positive preserved	0.27	86.1
Positive suppressed	0.04	13.9
New unsupported positive introduced	0.36	36.4
Positive recovered	0.04	3.7

Table 10: Paired bootstrap 95% confidence intervals for the RadFact logical-entailment evaluation on PadChest-GR. The table shows the absolute change (LOCI minus standard decoding). A positive ΔLF1 and a negative ΔUnsup/report favor LOCI.

Backbone	**ΔLPrec [95% CI]**	**ΔLRec [95% CI]**	**ΔLF1 [95% CI]**	**ΔUnsup/report [95% CI]**
LLaVA-Med 7B	$+2.3\ [0.6, 4.2]$	$-0.2\ [-1.6, 1.0]$	$+1.1\ [0.2, 2.2]$	$-0.17\ [-0.29, -0.03]$
Hulu-Med 7B	$+1.7\ [-0.2, 3.4]$	$-0.2\ [-1.9, 1.3]$	$+0.8\ [-0.4, 2.2]$	$-0.15\ [-0.31, 0.03]$
Qwen2.5-VL 7B	$+4.5\ [2.9, 6.3]$	$+0.0\ [-1.4, 1.6]$	$+2.3\ [1.2, 3.6]$	$-0.29\ [-0.46, -0.12]$

Table 11: Error-side coverage of the LOCI trigger on PadChest-GR. We replay standard-decoding reports through the frozen LOCI phrase matcher and scope gate and align triggered mentions with the official RadFact protocol used for the logical-entailment evaluation in Section 4.4. **FP coverage** is the percentage of standard-decoding unsupported positive findings exposed to a positive-scope LOCI trigger. **TP exposure** is the percentage of standard-decoding grounded positive findings exposed to such a trigger. **Correction@covered** is the percentage of covered unsupported positives that are corrected by LOCI, and **Suppression@exposed** is the percentage of exposed grounded positives that are suppressed by LOCI. The analysis is post-hoc and is not used for model or hyperparameter selection.

Backbone	Std. UnsupPos/report	FP coverage (%)	TP exposure (%)	Correction@covered (%)	Suppression@exposed (%)
LLaVA-Med 7B	1.35	60.8	70.4	39.8	13.5
Hulu-Med 7B	1.40	54.9	65.7	28.1	14.8
Qwen2.5-VL 7B	1.28	71.5	79.1	64.2	15.1

Table 12: Spatial alignment of concept directions with PadChest-GR radiologist bounding boxes. Pointing@1 is the percentage of grounded finding instances for which the maximally aligned image token has its receptive-field center inside an annotated box. BoxContrast is the mean within-image z-scored projection inside the box union minus the mean outside. Values are means with 95% study-clustered bootstrap confidence intervals. Area chance is the average fraction of valid image-token centers covered by the annotated boxes.

Backbone	Direction	Pointing@1 (%)	BoxContrast
LLaVA-Med 7B	Area chance	16.4 [14.8, 18.1]	–
	Random isotropic	16.8 [14.9, 18.3]	0.01 [−0.03, 0.05]
	Wrong concept	19.3 [17.2, 21.9]	0.06 [0.02, 0.08]
	Input embedding	21.7 [19.2, 24.3]	0.13 [0.08, 0.20]
	LM-head	23.4 [20.9, 27.1]	0.17 [0.11, 0.31]
	Same-layer matched (LOCI)	**29.6 [26.6, 32.9]**	**0.32 [0.24, 0.41]**
Hulu-Med 7B	Area chance	16.1 [14.5, 17.8]	–
	Random isotropic	16.4 [14.6, 18.4]	0.00 [−0.04, 0.07]
	Wrong concept	18.6 [16.5, 20.9]	0.04 [0.00, 0.12]
	Input embedding	20.5 [18.2, 23.0]	0.10 [0.05, 0.19]
	LM-head	21.6 [19.1, 24.2]	0.11 [0.07, 0.15]
	Same-layer matched (LOCI)	**27.2 [24.6, 30.0]**	**0.26 [0.19, 0.33]**
Qwen2.5-VL 7B	Area chance	15.8 [14.2, 17.5]	–
	Random isotropic	16.2 [14.4, 18.3]	0.01 [−0.03, 0.05]
	Wrong concept	21.0 [18.6, 23.5]	0.08 [0.03, 0.17]
	Input embedding	24.5 [21.9, 27.2]	0.17 [0.11, 0.29]
	LM-head	26.3 [23.6, 29.1]	0.22 [0.15, 0.29]
	Same-layer matched (LOCI)	**34.8 [31.7, 38.1]**	**0.43 [0.35, 0.57]**

A.2 Mechanistic Diagnostics and Ablation Studies

This section contains the full results for the mechanistic diagnostics and ablation studies discussed in Section 4.5, which investigate the components of the LOCI method.

Table 13: Sensitivity to the definition of the counterfactual margin on MIMIC-CXR with Qwen2.5-VL. All variants use the same phrase trigger, scope gate, same-layer concept directions, intervention layer, erasure strength, candidate set, group budget, and adaptive plausibility constraint. Only the form of the counterfactual margin is changed.

Decoding score	R-F1	14c-F1	5c-F1
Standard decoding	9.85	11.33	17.55
Normalized log-probability (LOCI)	11.07	13.53	22.88
Raw-logit difference	10.01	12.03	19.41
Top-K centered-logit difference	10.59	13.11	21.50

Table 14: Do text-derived prototypes carry image-side information? For each canonical finding mapped to a CheXbert observation, we project image-token hidden states at layer ℓ onto the candidate direction and measure how well the pooled projection separates reference-positive from reference-non-positive studies on the MIMIC-CXR validation split. *Matched* uses $\hat{u}_\phi$ to predict ϕ; *mismatched* uses $\hat{u}_\psi$ for $\psi \neq \phi$. Energy is the mean fraction of image-token norm removed at $\lambda = 1$. 5c-F1 is the corresponding test-set result from Table 5.

	Probe AUC			
Direction	Matched	Mismatched	Energy (%)	5c-F1
Concept-agnostic controls				
Random (isotropic)	0.50	n/a	1.30	17.45
Concept directions				
Input-embedding	0.56	0.52	3.35	19.34
LM-head	0.58	0.53	4.25	20.15
Same-layer text prototype (LOCI)	0.66	0.55	6.40	22.88

Table 15: Energy-matched control directions on MIMIC-CXR with Qwen2.5-VL. For each triggered event, the control direction is rescaled so that the removed image-token energy matches the LOCI matched-direction intervention at the same step. This isolates concept specificity from intervention magnitude.

Direction	Energy (%)	R-F1	14c-F1	5c-F1	Pos-Prec
Matched same-layer (LOCI)	6.40	**11.07**	**13.53**	**22.88**	**24.16**
Random isotropic (energy-matched)	6.38	10.06	11.88	18.31	17.24
Wrong concept (energy-matched)	6.42	9.97	11.74	18.08	16.91
LM-head (energy-matched)	6.36	10.29	12.46	20.11	19.85

Appendix B describes model loading, generation settings, prompts, dataset preprocessing, evaluator configuration, hyperparameter selection, and bootstrap uncertainty estimation. Appendix C gives the LOCI lexicon, phrase-matching procedure, and online scope gate. Appendix A.1 summarizes the grounded PadChest-GR evaluation protocol and grounded error analyses. Appendix C.4 and Appendix C.5 define the diagnostic controls and logging protocol. Appendix D reports supplementary analyses.

Unless explicitly stated otherwise, all prompts, lexicons, prototype banks, hyperparameters, and control settings are fixed using validation data before test evaluation. Test references and PadChest-GR reference annotations are not used for prompt design, lexicon construction, prototype construction, hyperparameter selection, or model selection.

B Implementation details

This section describes the implementation details for the report-generation experiments: hardware and model loading, generation settings, prompts, dataset preprocessing, evaluator configuration, LOCI hook locations, hyperparameter selection, and bootstrap uncertainty estimation. Details specific to the LOCI lexicon and scope gate are provided in Appendix C.

B.1 Hardware, model loading, and generation protocol

Hardware and model loading. Unless otherwise stated, experiments are run with batch size 1 on a single NVIDIA RTX 4090 GPU with 24GB VRAM. All LVLM checkpoints are used in frozen mode and loaded in float16 precision. We use the official or model-native image preprocessing code for each backbone. For decoding baselines that require attention weights, we disable attention kernels that do not expose the required attention tensors.

Generation protocol. All report-generation methods use the same model-specific chat template, image preprocessing pipeline, prompt semantics, stopping tokens, maximum generation length,

Table 16: Controlled input perturbation on the MIMIC-CXR single-view subset with Qwen2.5-VL. We compare LOCI against standard decoding under three input conditions: the true radiograph, a constant blank image, and a shuffled radiograph from another test study. All values are LOCI minus standard decoding. Pos-Prec and Pos-Rec are computed from CheXbert-5 labels by treating POS as positive and merging NEG/UNC/BLANK as non-positive. POS/report is the average number of CheXbert-positive findings per generated report.

Input to LVLM	ΔR-F1	Δ14c-F1	Δ5c-F1	ΔPos-Prec	ΔPos-Rec	ΔPOS/report	ΔLen
True image	+1.35	+2.40	+5.80	+8.42	+2.51	-0.23	-0.71
Blank image	+0.25	+0.50	+0.85	+1.51	-0.60	-0.06	-0.54
Shuffled image	+0.45	+0.90	+1.45	+2.41	-0.12	-0.06	-0.62

Table 17: Intervention-scope ablation on MIMIC-CXR with Qwen2.5-VL. Only the edited token set varies: the phrase trigger, scope gate, concept direction, erasure scale λ, hook layer, and reranking rule are held fixed, and the input image is the true radiograph in all rows. The three conditions correspond to $\mathbf{m} = m^{\text{vis}}$, $\mathbf{m} = \mathbf{1} - m^{\text{vis}}$, and $\mathbf{m} = \mathbf{1}$ in Eq. (1). Energy is the mean fraction of hidden-state norm removed at edited positions, averaged over triggered steps. Pos-Prec and Pos-Rec use the CheXbert-5 binarization of Section 4.3.

Edited positions	Energy (%)	R-F1	14c-F1	5c-F1	Pos-Prec	Pos-Rec
None (standard decoding)	0.00	9.85	11.33	17.55	16.34	20.61
Text tokens only	8.85	10.07	11.72	18.31	17.02	19.98
All tokens	6.72	10.63	12.66	20.89	20.72	21.40
Image tokens only (LOCI)	6.40	**11.07**	**13.53**	**22.88**	**24.16**	**22.40**

and repetition penalty. Unless otherwise noted, generation is deterministic greedy decoding with `do_sample=False`, `max_new_tokens=150`, and repetition penalty $r_{\text{rep}} = 1.2$. These shared settings are fixed before comparing decoding methods and are not tuned for LOCI.

Baselines. We compare LOCI with standard greedy decoding and five training-free decoding baselines: VCD, Med-VCD, OPERA, ProjectAway, and VTI. All methods use the same prompts, image preprocessing, stopping rules, maximum generation length, and repetition penalty. When official implementations are available, we use them; otherwise, we reproduce the published decoding rule. Method-specific hyperparameters are selected on validation data disjoint from all test sets.

Evaluation overview. Final generated reports are evaluated using the pipeline in Appendix B.4. The LOCI lexicon and scope-gating rules are not applied to evaluator inputs. ROUGE-L is reported only as a lexical-overlap diagnostic.

Baseline-specific implementation details. To make the comparison reproducible, we explicitly distinguish shared report-generation settings from method-defining baseline components. All baselines use the same report-generation instruction semantics, image preprocessing, stopping criteria, and maximum generation length as LOCI. When a method is defined by a native auxiliary branch or search policy, we preserve that component and tune only the published method-specific hyperparameters on the MIMIC-CXR validation split before freezing them for all test datasets.

B.2 Prompts and prototype extraction

All report-generation experiments use the same instruction semantics across backbones. Only the model-native chat wrapper and image block format differ across checkpoints. The following prompts are the exact prompt semantics used in the reported experiments.

Shared report-generation instruction. The following instruction text is used for all chest X-ray report-generation experiments:

```
You are a radiology assistant. Given the chest X-ray image, write a
↪  concise
chest X-ray report in standard radiology style. Describe only findings
```

Table 18: Incremental predictive value of the LOCI counterfactual margin on MIMIC-CXR with Qwen2.5-VL. All input conditions use the same fixed candidate events and prefixes. Logistic coefficients and finding-specific intercepts are fitted on validation data and frozen before test evaluation. Each cell reports test AUC with a 95% study-clustered bootstrap interval. Incremental ΔAUC is AUC(Base + Margin) minus AUC(Base).

Input signal	**True image**	**Shuffled image**	**Blank image**
Base score only	$0.61\ [0.58, 0.63]$	$0.53\ [0.50, 0.55]$	$0.52\ [0.50, 0.54]$
Margin only	$0.62\ [0.59, 0.65]$	$0.52\ [0.49, 0.54]$	$0.51\ [0.48, 0.54]$
Base score + margin	$0.68\ [0.65, 0.71]$	$0.53\ [0.51, 0.56]$	$0.53\ [0.50, 0.55]$
Incremental ΔAUC	$+0.07\ [0.05, 0.09]$	$+0.00\ [-0.01, 0.02]$	$+0.01\ [0.00, 0.02]$

Table 19: Low-cost projection-based reranking controls on MIMIC-CXR with Qwen2.5-VL. These controls avoid a full counterfactual forward pass and score matched candidates using only image-token alignment with the matched same-layer prototype.

Method	R-F1	14c-F1	5c-F1	Pos-Prec	Relative latency
Max projection reranker	10.21	12.35	19.94	18.76	1.14×
Mean projection reranker	10.34	12.52	20.27	19.11	1.16×
Projection-only contrastive score	10.46	12.63	20.71	19.84	1.19×
Scope-aware direct erasure	10.88	12.77	17.98	17.48	2.89×
LOCI	**11.07**	**13.53**	**22.88**	**24.16**	2.90×

```
supported by the image. If a finding is uncertain, use uncertain wording
rather than a definite assertion. Output only the report text.
```

LLaVA-Med v1.5. For the LLaVA-style backbones, we use a single-turn chat wrapper with one image placeholder and one user instruction:

```
USER: <image>
You are a radiology assistant. Given the chest X-ray image, write a
↪  concise
chest X-ray report in standard radiology style. Describe only findings
supported by the image. If a finding is uncertain, use uncertain wording
rather than a definite assertion. Output only the report text.
ASSISTANT:
```

Hulu-Med-7B. It uses its own ChatML template with the model's built-in system prompt (`add_system_prompt=True`):

```
<|im_start|>system
You are Hulu-Med, a helpful health assistant that can understand text, 2D
↪  images,
videos, and 3D images.<|im_end|>
<|im_start|>user
<image>
{instruction}<|im_end|>
<|im_start|>assistant
```

Qwen2.5-VL-7B-Instruct. For Qwen2.5-VL, we use the model-native chat template with one system message and one user message containing both an image block and a text block:

```
[
  {"role": "system", "content": "You are a helpful radiology
  ↪  assistant."},
  {"role": "user",
   "content": [
     {"type": "image", "image": "<image>"},
     {"type": "text",
      "text": "Given the chest X-ray image, write a concise chest X-ray
      ↪  report
```

Table 20: No-op full-forward identity check. For identical image–prefix pairs, we compare cached base logits with logits from the complete counterfactual full-forward path at $\lambda = 0$. Mean and maximum differences are computed over finite raw full-vocabulary logits before token masking. Top-1 agreement is computed after applying identical deterministic decoding processors. Exact report agreement is token-sequence equality between standard decoding and LOCI with $\lambda = 0$.

Backbone	**Reports**	**Prefixes**	**Mean $\lvert\Delta z\rvert$**	**Max $\lvert\Delta z\rvert$**	**Top-1 agree. (%)**	**Exact reports (%)**
LLaVA-Med 7B	300	35,580	1.70×10^{-5}	2.10×10^{-3}	100.00	100.00
Hulu-Med 7B	300	30,420	1.50×10^{-5}	1.90×10^{-3}	100.00	100.00
Qwen2.5-VL 7B	300	37,410	2.40×10^{-5}	3.20×10^{-3}	100.00	100.00

Table 21: Backbone-specific LOCI implementation details. Layer indices are 0-indexed. The hyperparameters are selected on the validation split using CheXbert-5 F1, with RadGraph F1 used as a tie-breaker. Search grids for LOCI and the clinical controls are reported in Table 28.

Backbone	Hook module path	Hook point	ℓ	K_{cand}	$G_{\max}$	$(\alpha, \lambda, \alpha_{\text{apc}})$
LLaVA-Med 7B	`language_model.model.layers`	post-block residual	24	30	3	(0.60, 1.0, 0.3)
Hulu-Med 7B	`language_model.model.layers`	post-block residual	20	30	3	(0.40, 1.0, 0.3)
Qwen2.5-VL 7B	`model.language_model.layers`	post-block residual	18	30	3	(0.80, 1.0, 0.3)

```
in standard radiology style. Describe only findings supported by the
↪  image.
If a finding is uncertain, use uncertain wording rather than a definite
assertion. Output only the report text."}
   ]}
]
```

Prototype-extraction prompt. For same-layer prototype extraction, we use the same backbone-specific chat template as above, but remove the image block and set the assistant continuation to the canonical finding phrase itself. For example, for the canonical finding *pleural effusion*, the assistant target is the literal string `pleural effusion`. This ensures that prototype extraction and hidden-state intervention are performed in the same prompt-conditioned representation space.

Decoding parameters shared across prompts. Unless otherwise noted, all report-generation runs use deterministic greedy decoding with `do_sample=False`, `max_new_tokens=150`, and `repetition_penalty=1.2`. The same stopping criteria are used for all methods compared in the main text.

Optional conservative-prompt control. We append the following sentence to the shared instruction text and keep all other settings unchanged:

```
If the visual evidence is weak, prefer uncertain wording or omit the
↪  finding
rather than making a definite unsupported assertion.
```

B.3 Dataset Splits and Report Preprocessing

For study-based datasets, we perform splitting at the study level before image-level expansion, so that all views from the same study remain in the same split. After split assignment, each image in a multi-view study is evaluated independently against the corresponding study report.

Study-level splitting and image-level expansion. Whenever a report corresponds to more than one image, we first assign the *study* to a split and only then expand that study to one row per image. Thus, all views from a given study remain in the same split. The main-paper protocol is image-level: after split assignment, each image in a multi-view study is evaluated independently against the same study report.

Section extraction. We extract the reference report from the *Findings* and *Impression* sections when both are available. If a dataset provides structured section columns, we use them directly. Otherwise,

we apply case-insensitive heading matching to the raw report using the section names `findings`, `finding`, `impression`, and `conclusion`, with `conclusion` mapped to `impression`. If both extracted sections are non-empty, the final reference is

$$\texttt{reference_text} = \texttt{findings_text} \,||\, \texttt{"\textbackslash n"} \,||\, \texttt{impression_text}.$$

If only one section is present, we use the available section alone. If neither section can be recovered after cleanup, the example is dropped.

Text cleanup. The following preprocessing is applied to all references before evaluation:

1. Unicode normalize to NFKC.
2. Remove duplicated section headers if they remain in the extracted text.
3. Collapse repeated whitespace within lines.
4. Collapse repeated blank lines to a single newline.
5. Strip leading and trailing whitespace.
6. Preserve original case and punctuation except for the cleanup steps above.

No lexical synonym mapping, lexicon normalization, or cue-word normalization is applied to the evaluator inputs.

Dataset-specific notes.

- **MIMIC-CXR.** The official patient/study split is respected at the study level.
- **IU-Xray.** Each report may be associated with one or more images.
- **CheXpert Plus.** We use the report-bearing subset.
- **PadChest-GR.** We use the report-bearing subset.

B.4 Evaluators and label normalization

Table 22 summarizes the evaluator configuration used in the main paper.

Table 22: Evaluator configuration and label normalization.

Metric	Software / version	Input normalization	Output handling
RadGraph F1	`v0.1.18`	Unicode NFKC; collapse repeated whitespace; no lexicon normalization	Official entity–relation F1 on generated versus reference report text
CheXbert 14c-F1 / 5c-F1	`v0.0.2`	Same as above	CheXbert labels in {POS, NEG, UNC, BLANK}; macro-F1 over 14 or 5 observations
Positive precision / recall	Same CheXbert outputs as above	Same as above	POS treated as positive; NEG, UNC, and BLANK merged into NONPOS; macro-averaged over 5 observations
ROUGE-L	`2022-07`	Same cleaned report text	Lexical-overlap diagnostic only
RadFact logical P/R	`main branch`	Generated and reference reports are converted to finding-level sentences using `GPT-4`	Logical entailment evaluated with `Llama-3-70B-Instruct`; LPrec is generated→reference entailment, LRec is reference→generated entailment; LF1 is their harmonic mean; no spatial boxes are used

CheXbert observations. For 14-category CheXbert evaluation, we use the following observations: *Enlarged Cardiomediastinum*, *Cardiomegaly*, *Lung Opacity*, *Lung Lesion*, *Edema*, *Consolidation*, *Pneumonia*, *Atelectasis*, *Pneumothorax*, *Pleural Effusion*, *Pleural Other*, *Fracture*, *Support Devices*, and *No Finding*. The 5-category subset used in the main text contains *Atelectasis*, *Cardiomegaly*, *Consolidation*, *Edema*, and *Pleural Effusion*.

Label normalization. Different CheXbert implementations may emit labels as strings or numeric codes. Before scoring, we normalize them to a common label set:

$$\{\text{POS}, \text{NEG}, \text{UNC}, \text{BLANK}\}.$$

If the implementation emits numeric states, we map them as

$$1 \to \text{POS}, \qquad 0 \to \text{NEG}, \qquad -1 \to \text{UNC},$$

and treat missing / no-mention outputs as BLANK.

CheXbert positive-class evaluation. For each observation o, we binarize both the generated-report and reference-report labels by treating POS as the positive class and merging NEG, UNC, and BLANK into NONPOS. Let P_o, R_o, and $F_{1,o}$ denote the resulting positive-class precision, recall, and F1 for observation o. Zero-denominator cases are assigned a score of zero.

For an observation set $\mathcal{O}$, we report

$$\text{Macro-P} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} P_o, \qquad \text{Macro-R} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} R_o,$$

and

$$\text{Macro-F1} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} F_{1,o}.$$

We use $|\mathcal{O}| = 14$ for 14c-F1 and $|\mathcal{O}| = 5$ for 5c-F1. Throughout the paper, 14c-F1 and 5c-F1 refer to the macro-average of the per-observation positive-class F1 scores.

Generated-report assertion statistics. For the descriptive assertion statistics, let $\hat{y}_{i,o} \in \{\text{POS}, \text{NEG}, \text{UNC}, \text{BLANK}\}$ denote the CheXbert label assigned to observation o in generated report i.

For assertion state $s \in \{\text{POS}, \text{NEG}, \text{UNC}\}$, we define

$$\text{State/report}(s) = \frac{1}{N} \sum_{i=1}^{N} \sum_{o \in \mathcal{O}_5} \mathbf{1}[\hat{y}_{i,o} = s],$$

where $\mathcal{O}_5$ is the five-observation subset used for 5c-F1.

These statistics characterize the assertion distribution of generated reports and do not compare the generated reports with references. The CheXbert evaluator is applied directly to the generated text and does not use the LOCI phrase lexicon or scope gate.

Baseline-to-LOCI edit analysis. Let y_o^{ref}, y_o^{base}, and y_o^{loci} be the binarized reference, standard-decoding, and LOCI labels for observation o. We report:

$$\text{FP-corr} = \frac{\#\{o : y_o^{\text{base}} = 1,\ y_o^{\text{ref}} = 0,\ y_o^{\text{loci}} = 0\}}{\#\{o : y_o^{\text{base}} = 1,\ y_o^{\text{ref}} = 0\}},$$

$$\text{TP-supp} = \frac{\#\{o : y_o^{\text{base}} = 1,\ y_o^{\text{ref}} = 1,\ y_o^{\text{loci}} = 0\}}{\#\{o : y_o^{\text{base}} = 1,\ y_o^{\text{ref}} = 1\}}.$$

These rates are macro-averaged over the same five observations as the 5c-F1 metric.

RadGraph evaluation. RadGraph is run directly on the final cleaned report strings. We do not apply the LOCI lexicon, phrase normalization, or scope cue normalization to evaluator inputs. If the evaluator returns separate entity and relation scores, we report the combined official F1 returned by the evaluation script.

Paired clustered bootstrap intervals. We use 10,000 paired percentile-bootstrap resamples. For MIMIC-CXR, CheXpert Plus, and PadChest-GR, the resampling unit is the study: a sampled study contributes all of its evaluated image–report pairs. For IU-Xray, which can associate multiple images with a single report, we cluster resampling by the report unique identifier (UID) to account for non-independent samples. The same resampled units are used for all methods within a backbone–dataset condition, and each metric is recomputed from the complete resampled prediction set.

Table 23: Paired bootstrap 95% confidence intervals for LOCI minus standard decoding. Intervals use 10,000 paired bootstrap resamples, clustered by study for MIMIC-CXR, CheXpert Plus, and PadChest-GR; IU-Xray is resampled at the released report/image-unit level.

Backbone	Dataset	ΔR-F1	Δ14c-F1	Δ5c-F1
LLaVA-Med 7B	MIMIC-CXR	$+1.43\ [0.69, 2.18]$	$+0.28\ [-0.21, 1.04]$	$+1.93\ [0.95, 2.92]$
LLaVA-Med 7B	IU-Xray	$+1.68\ [0.09, 3.29]$	$+1.34\ [0.18, 2.51]$	$+2.12\ [0.30, 3.96]$
LLaVA-Med 7B	CheXpert Plus	$+0.53\ [-1.45, 2.81]$	$+0.28\ [-2.02, 2.29]$	$+3.84\ [0.28, 8.19]$
LLaVA-Med 7B	PadChest-GR	$+0.07\ [-0.25, 0.39]$	$+0.41\ [-1.61, 2.42]$	$+0.24\ [-1.69, 2.31]$
Hulu-Med 7B	MIMIC-CXR	$+0.44\ [0.06, 0.83]$	$+0.46\ [0.09, 0.84]$	$+0.68\ [-0.29, 1.65]$
Hulu-Med 7B	IU-Xray	$+2.74\ [0.85, 4.65]$	$+0.72\ [-0.99, 2.41]$	$+3.18\ [0.99, 5.39]$
Hulu-Med 7B	CheXpert Plus	$+0.18\ [-3.08, 3.65]$	$+1.06\ [-1.66, 4.05]$	$+2.48\ [-1.05, 6.28]$
Hulu-Med 7B	PadChest-GR	$+0.04\ [-0.41, 0.49]$	$+0.70\ [-1.25, 2.66]$	$+0.62\ [-1.61, 3.02]$
Qwen2.5-VL 7B	MIMIC-CXR	$+1.22\ [0.52, 1.91]$	$+2.20\ [1.40, 2.99]$	$+5.33\ [4.35, 6.32]$
Qwen2.5-VL 7B	IU-Xray	$+0.27\ [-1.69, 2.24]$	$+0.38\ [-1.38, 2.14]$	$+0.30\ [-1.79, 2.54]$
Qwen2.5-VL 7B	CheXpert Plus	$+0.30\ [-3.26, 3.61]$	$+1.66\ [-1.14, 4.79]$	$+3.15\ [-0.49, 7.06]$
Qwen2.5-VL 7B	PadChest-GR	$+3.25\ [1.78, 4.74]$	$+1.54\ [-0.40, 3.49]$	$+4.11\ [1.97, 6.26]$

B.5 LOCI HOOK LOCATIONS AND HYPERPARAMETERS

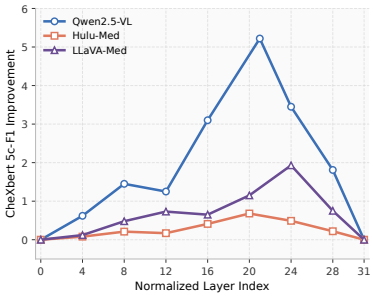


Figure 6: Layer-wise intervention analysis on the MIMIC-CXR validation split. We sweep the decoder layer used for concept erasure and report the absolute improvement in CheXbert 5c-F1 over standard decoding.

For each backbone, prototype extraction and counterfactual editing are performed at the same post-block residual stream: the tensor returned by decoder block ℓ after self-attention, MLP, and residual updates, before the input normalization of block $\ell + 1$. The intervention layer and LOCI hyperparameters are selected on validation data and then fixed for test evaluation. Table 21 reports the selected configuration.

B.6 UNCERTAINTY ESTIMATION

We report paired bootstrap confidence intervals for LOCI minus standard decoding. For datasets with study identifiers (MIMIC-CXR, CheXpert Plus, and PadChest-GR), we perform study-level clustered bootstrap. For IU-Xray, which can associate multiple images with a single report, we cluster resampling by the report unique identifier (UID) to account for non-independent samples. The same resampled units are used for all methods within a backbone–dataset condition.

Comparison with the strongest competing decoder. Reporting only LOCI-minus-standard uncertainty does not establish whether LOCI is statistically distinguishable from the strongest alternative training-free decoder. We therefore perform an additional conservative paired comparison against the competing method that performs best within each bootstrap resample.

Let $\mathcal{B} = \{\mathrm{VCD}, \mathrm{Med} - \mathrm{VCD}, \mathrm{OPERA}, \mathrm{ProjectAway}, \mathrm{VTI}\}$ denote the set of competing training-free decoders. For bootstrap resample r and metric M, define

$$\Delta^{(r)}_{\mathrm{best}} = M^{(r)}_{\mathrm{LOCI}} - \max_{b \in \mathcal{B}} M^{(r)}_{b}.$$

All methods are evaluated on exactly the same resampled studies or image/report units. We report the 2.5th and 97.5th percentiles of $\{\Delta^{(r)}_{\mathrm{best}}\}_{r=1}^{10000}$ as the paired 95% confidence interval.

Unlike a pairwise comparison with a single test-set runner-up, this procedure allows the identity of the strongest competing method to vary across bootstrap resamples. It therefore avoids conditioning the uncertainty analysis on a competitor selected from the observed test scores and provides a conservative test of whether LOCI exceeds the strongest available competing decoder. Note that the full-test-set point estimate is not mathematically required to fall within this percentile-based interval due to the selection bias introduced by the max operator at each resample.

Table 24: Paired comparison of LOCI with the strongest competing training-free decoder. At each bootstrap resample, the competing score is the maximum over VCD, Med-VCD, OPERA, ProjectAway, and VTI evaluated on the same resampled units. Positive values favor LOCI.

Backbone	Dataset	Δ_{best} R-F1	95% CI	Δ_{best} 5c-F1	95% CI
LLaVA-Med 7B	MIMIC-CXR	+1.26	$[+0.47, +2.02]$	+4.95	$[+3.76, +6.10]$
	IU-Xray	+0.99	$[+0.05, +2.09]$	-0.14	$[-2.03, +1.75]$
	CheXpert Plus	+0.58	$[-1.54, +2.75]$	+2.23	$[-1.12, +5.53]$
	PadChest-GR	+0.04	$[-1.16, +1.22]$	+0.17	$[-1.58, +1.92]$
Hulu-Med 7B	MIMIC-CXR	+0.49	$[-0.05, +1.03]$	+1.43	$[+0.81, +2.05]$
	IU-Xray	+3.39	$[+1.21, +5.57]$	+3.30	$[+1.52, +5.08]$
	CheXpert Plus	+0.41	$[-2.10, +2.44]$	+2.62	$[-2.35, +4.55]$
	PadChest-GR	+0.11	$[-1.05, +1.25]$	+0.89	$[-1.25, +2.16]$
Qwen2.5-VL 7B	MIMIC-CXR	+0.72	$[+0.18, +1.27]$	+1.38	$[+0.50, +2.21]$
	IU-Xray	+1.47	$[-0.45, +3.15]$	+0.05	$[-2.25, +1.98]$
	CheXpert Plus	+0.48	$[-2.34, +3.28]$	+1.12	$[-2.60, +4.80]$
	PadChest-GR	+2.76	$[+1.26, +4.21]$	+0.53	$[-1.98, +3.04]$

C Clinical Lexicon, Phrase Matching, and Scope Gate

C.1 Clinical Phrase Lexicon

LOCI uses a fixed phrase-level clinical lexicon and an online left-context scope gate to decide when a candidate continuation completes a positive-scope clinical finding phrase. The lexicon is used only for LOCI triggering, same-concept grouping, and diagnostic controls that explicitly reuse the LOCI trigger. It is not used by RadGraph, CheXbert, ROUGE-L, or the PadChest-GR RadFact-based evaluator.

Normalization used before matching. All aliases are normalized with the same function $\mathcal{N}$ used at test time: lowercasing, Unicode normalization, collapsing repeated whitespace, mapping hyphen variants to spaces, and stripping only trailing punctuation for matching (`. , ; : ) ]`). This normalization is applied for trigger matching only; it is not applied to the evaluator inputs.

Optional preposed modifiers. A matched alias may be preceded by one modifier chunk from the following set:

> `left`, `right`, `bilateral`, `upper lobe`, `lower lobe`, `middle lobe`, `upper lung`, `lower lung`, `apical`, `biapical`, `basilar`, `bibasilar`, `perihilar`, `retrocardiac`, `hilar`.

Modifiers are stripped when mapping a matched surface span to its canonical finding. For example, *left pleural effusion* maps to the canonical finding *pleural effusion*. Post-nominal anatomy phrases such as *pleural effusion in the right hemithorax* are not treated as modifiers for trigger construction.

Canonical findings and matched surface forms.

- **pleural effusion**: pleural effusion; effusion; pleural fluid.
- **pneumothorax**: pneumothorax; ptx; apical pneumothorax.
- **focal consolidation**: focal consolidation; consolidation; airspace consolidation.
- **lung opacity**: opacity; lung opacity; airspace opacity; pulmonary opacity.
- **pulmonary edema**: pulmonary edema; edema; interstitial edema; alveolar edema.
- **atelectasis**: atelectasis; subsegmental atelectasis; bibasilar atelectasis; discoid atelectasis.
- **cardiomegaly**: cardiomegaly; enlarged cardiac silhouette; enlarged heart; cardiac enlargement.
- **enlarged cardiomediastinal silhouette**: enlarged cardiomediastinal silhouette; widened cardiomediastinal silhouette; enlarged cardiomediastinum.
- **pulmonary nodule**: pulmonary nodule; lung nodule; nodular opacity; nodular density; nodule.
- **pulmonary mass**: pulmonary mass; lung mass; mass lesion; hilar mass.

Algorithm 1 Online phrase matching and scope gating

Input: detokenized prefix $y_{<t}$, top-K candidate set $\mathcal{C}_t$, alias bank $\mathcal{A}$, modifier set $\mathcal{M}_{\text{mod}}$, cue sets $\mathcal{P}_{\text{pos}}$, $\mathcal{P}_{\text{unc}}$, $\mathcal{P}_{\text{neg}}$
initialize empty finding groups $\mathcal{G}_t$
for all $c \in \mathcal{C}_t$ **do**
  $y_{\text{cand}} \leftarrow \text{TokDec}([c])$ {Detokenize candidate token}
  $y_{\text{full}} \leftarrow y_{<t} \oplus y_{\text{cand}}$ {Append to prefix}
  $z \leftarrow \text{CurrentClause}(\mathcal{N}(y_{\text{full}}))$ {Normalize and extract current clause}
  $(\phi, b, e) \leftarrow \text{LongestSuffixMatch}(z, \mathcal{A}, \mathcal{M}_{\text{mod}})$
  if $(\phi, b, e) = \varnothing$ **then**
    continue
  end if
  $p \leftarrow z[1 : b-1]$ {Left context before the matched finding}
  $\sigma \leftarrow \text{ScopeLabel}(p, \mathcal{P}_{\text{pos}}, \mathcal{P}_{\text{unc}}, \mathcal{P}_{\text{neg}})$
  if $\sigma = \text{POS}$ **then**
    add c to group $\mathcal{M}_t(\phi)$
  end if
end for
sort non-empty groups by the best base logit among their members
keep at most $G_{\max}$ groups
return $\mathcal{G}_t = \{(\phi, \mathcal{M}_t(\phi)) : \mathcal{M}_t(\phi) \neq \varnothing\}$

- **lung lesion**: lung lesion; pulmonary lesion; parenchymal lesion.
- **infiltrate**: infiltrate; infiltrative opacity; pulmonary infiltrate.
- **pneumonia**: pneumonia; multifocal pneumonia; lobar pneumonia.
- **pleural thickening**: pleural thickening; apical pleural thickening; biapical pleural thickening; pleural apical thickening.
- **pleural plaque**: pleural plaque; pleural plaques; calcified pleural plaque.
- **fibrosis**: fibrosis; fibrotic change; fibrotic opacity.
- **scarring**: scarring; scar; linear scarring.
- **hyperinflation**: hyperinflation; hyperinflated lungs; hyperexpanded lungs.
- **emphysema**: emphysema; emphysematous change; emphysematous lungs.
- **interstitial opacity**: interstitial opacity; interstitial infiltrate; reticular opacity; diffuse interstitial opacity.
- **hilar prominence**: hilar prominence; hilar enlargement; prominent hila.
- **mediastinal widening**: widened mediastinum; mediastinal widening; enlarged mediastinum.
- **fracture**: fracture; acute fracture; displaced fracture.
- **hiatal hernia**: hiatal hernia; large hiatal hernia; paraesophageal hernia.
- **aspiration**: aspiration; aspiration opacity; aspiration pneumonia.
- **support device**: support device; support devices; lines and tubes; medical device.

Notes on ambiguity. When several aliases match at the current step, LOCI selects the longest normalized suffix, with ties broken by character length. Thus *pleural effusion* overrides *effusion*, and *apical pneumothorax* overrides *pneumothorax* when both are possible suffix matches. The machine-readable JSON released with the code is authoritative and should be used for exact reproduction.

C.2 Phrase matching and scope gating

Algorithm 1 gives the exact online trigger procedure used by LOCI. All matching is performed on detokenized text, not on raw subword IDs. The released implementation is included in `code/phrase_matcher.py` and `code/scope_gate.py`.

Definition of helper functions.

- $\mathcal{N}(\cdot)$: lowercase, Unicode-normalize, collapse repeated whitespace, map hyphen variants to spaces, and strip only trailing punctuation for matching.
- $\mathrm{CurrentClause}(\cdot)$: return the suffix after the most recent sentence boundary or adversative boundary. Sentence boundaries are `.` `?` `!` `;`; adversative boundaries include `but`, `however`, `although`, `whereas`, and `despite`. Commas and coordinators (`and`, `or`) do not reset scope because radiology reports frequently use coordinated negations, e.g., *no focal consolidation, edema, or pneumothorax*.
- $\mathrm{LongestSuffixMatch}(\cdot)$: check whether the current clause ends on a word boundary with an alias from the lexicon, optionally preceded by exactly one modifier chunk from $\mathcal{M}_{\mathrm{mod}}$. If several aliases match, choose the longest normalized suffix; break ties by character length.
- $\mathrm{ScopeLabel}(\cdot)$: apply precedence-ordered pattern matching to the normalized left context: POS-override/persistence $>$ UNC $>$ NEG $>$ default POS.

Cue sets used by the scope gate. The exact cue lists are:

- **POS-override/persistence cues:** `no change in`, `unchanged`, `stable`, `persistent`, `residual`, `decreased but persistent`.
- **UNC cues:** `possible`, `possibly`, `may represent`, `could represent`, `cannot exclude`, `difficult to exclude`, `questionable`.
- **NEG cues:** `no`, `without`, `free of`, `negative for`, `absence of`, `resolved`, `resolution of`.

Illustrative scope decisions. With the above rules, *no focal consolidation, edema, or pneumothorax* is NEG; *cannot exclude small effusion* is UNC; *no change in pleural effusion* is POS; *resolved pleural effusion* is NEG; and *decreased but persistent effusion* is POS.

Known online limitation. Because the trigger is computed before future tokens are generated, post-nominal cues that only appear after the finding span, such as *effusion is not excluded*, are not visible at trigger time and are therefore treated as a known failure mode rather than a preprocessing error.

C.3 Scope-gate validation

We evaluate the online scope gate as a text-only diagnostic. We extract report sentences containing at least one LOCI lexicon alias and compare the gate's POS/NEG/UNC decisions with NegBio-derived silver assertion labels. The gate observes only the left context available at phrase-completion time, matching its use during decoding. This diagnostic does not use images and is not used for model selection.

Tables 25 and 26 report per-label precision/recall/F1 and the confusion matrix. The gate performs well on common pre-nominal negation and uncertainty patterns, while post-nominal cues such as *effusion is not excluded* remain a known online limitation.

Table 25: Scope-gate validation against silver assertion labels on MIMIC-CXR reports. The LOCI scope gate observes only the left context available at phrase-completion time. Precision, recall, and F1 are computed for POS, NEG, and UNC labels.

Scope label	Precision	Recall	F1	Support
POS	91.2	94.8	93.0	42360
NEG	92.9	90.6	91.7	31240
UNC	78.7	66.9	72.3	5860

C.4 Clinical control implementation and tuning

The clinical controls in Table 2 are diagnostic controls rather than generic decoding baselines. They are designed to test whether LOCI's gains can be explained by simpler alternatives: conservative

Table 26: Scope-gate confusion matrix against silver assertion labels. Rows are silver labels and columns are LOCI scope-gate predictions.

Silver \ Predicted	POS	NEG	UNC
POS	40150	1550	660
NEG	2550	28290	400
UNC	1320	620	3920

prompting, fixed phrase-level penalties, generic blank-image contrast, or arbitrary image-token perturbations.

All clinical controls are tuned on the same validation split used for LOCI. Unless otherwise noted, we select the setting with the highest validation CheXbert-5 F1 and use validation RadGraph F1 as a tie-breaker. The selected configuration is then fixed before test evaluation. Test references are never used for prompt selection, hyperparameter selection, or model selection.

All controls use the same report-generation instruction format, chat template, image preprocessing, maximum generation length, stopping rules, and repetition penalty as standard decoding and LOCI. For controls that use phrase triggering, we use the same detokenized phrase matcher, scope gate, top-K candidate set, and maximum finding-group budget as LOCI.

Table 27: Conservative prompt variants used for validation selection. P_0 is the original report-generation prompt. We select the variant with the highest validation CheXbert-5 F1 and use it as the tuned conservative-prompt control.

ID	Added instruction
P_0	None; original report-generation instruction.
P_1	If the visual evidence is weak, prefer uncertain wording or omit the finding rather than making a definite assertion.
P_2	Only mention abnormalities that are clearly supported by the image. Do not over-call uncertain findings.
P_3	Avoid definite positive findings unless they are visually evident. Use "possible" or "may represent" for uncertain abnormalities.
P_4	Be conservative when describing abnormalities. It is better to omit an unsupported finding than to report it as definitely present.

Tuned conservative prompt. This control preserves the original report-generation task and changes only the instruction text. We evaluate the prompt variants in Table 27 on the validation split and select the one with the highest validation CheXbert-5 F1.

Scope-aware lexicon penalty. This control uses the same phrase trigger and scope gate as LOCI, but does not run a counterfactual branch and does not edit hidden states. Let U_t be the set of matched positive-scope finding candidates at decoding step t. For each $y \in U_t$, we update the base logit by

$$S_t(y) = \begin{cases} \log p_t^b(y) - \beta, & y \in \mathcal{U}_t, \\ \log p_t^b(y), & \text{otherwise.} \end{cases}$$

where β is selected on the validation split. All unmatched candidates keep their base logits. This control tests whether LOCI's gains can be reproduced by simply penalizing positive finding phrases.

Blank-image contrastive control. This control keeps LOCI's phrase trigger and scope gate, but replaces the matched concept-direction intervention with a generic input-level visual ablation. At a triggered step, the base branch uses the true image, while the auxiliary branch uses a constant blank image with the same text prefix. The auxiliary branch yields a next-token distribution p_t^{blank}. For matched candidates $y \in \mathcal{U}_t$, we compute

$$m_t^{\text{blank}}(y) = \log p_t^b(y) - \log p_t^{\text{blank}}(y),$$

and define

$$S_t^{\text{blank}}(y) = \begin{cases} \log p_t^b(y) + \alpha_{\text{blank}} m_t^{\text{blank}}(y), & y \in \mathcal{U}_t,\ a_t(y) = 1, \\ \log p_t^b(y), & \text{otherwise.} \end{cases}$$

This control tests whether a generic visual contrastive signal is sufficient without a matched clinical concept direction.

Random-direction intervention. This control uses the same phrase trigger, scope gate, hook layer, and image-token mask as LOCI. It applies a reranking rule of the same form as Eq. 2, but replaces the matched phrase direction $\hat{u}_\phi$ with a random normalized direction $\hat{r}_\phi$. For both Tables 2 and 5, this control fixes $\alpha_{\text{apc}} = 0.30$; α and λ are tuned independently on the validation set (see Table 28). For each canonical finding, we sample a random vector from a standard Gaussian distribution and normalize it to unit ℓ_2 norm. The random direction bank is fixed during decoding. This control tests whether LOCI's gains depend on concept-specific phrase directions rather than arbitrary image-token perturbations. We use a random direction bank across seeds.

Scope-aware direct-erasure control. This control isolates whether the gains arise from matched concept erasure itself or from using the erasure as a relative counterfactual signal. It uses exactly the same phrase trigger, scope gate, matched same-layer prototype, image-token mask, hook layer, candidate set, and group budget as LOCI.

For a matched candidate $y \in \mathcal{U}_t$, instead of rewarding sensitivity to the counterfactual intervention, the control moves the candidate score toward the counterfactual distribution:

$$S_t^{\text{direct}}(y) = (1-\eta)\log p_t^b(y) + \eta \log p_t^{\text{cf},\phi_y}(y), \qquad y \in \mathcal{U}_t,\ a_t(y) = 1,$$

and uses $\log p_t^b(y)$ otherwise. Thus, $\eta = 1$ corresponds to using the erased candidate score directly, whereas LOCI moves in the opposite direction by rewarding $\log p_t^b(y) - \log p_t^{\text{cf},\phi_y}(y)$.

We keep the LOCI-selected intervention layer and erasure strength fixed and select $\eta \in \{0.25, 0.50, 1.00\}$ on the same validation split. This control therefore separates matched concept removal from the relative counterfactual scoring rule proposed by LOCI.

C.5 Trigger logging and counterfactual margin diagnostics

For each generated report, we log the number of decoded tokens, phrase-match events, scope labels, processed finding groups, counterfactual forward passes, and final decision changes. A phrase-match event is counted when a top-K candidate completes a lexicon alias on a word boundary after detokenization. If multiple aliases match the same suffix, the longest alias is used, as in the main LOCI trigger. POS matches are processed by LOCI, while NEG and UNC matches are skipped. A decision change is counted when the token selected by LOCI differs from the token that would have been selected by standard greedy decoding under the same prefix.

No-op full-forward identity diagnostic. The standard branch uses incremental KV-cache decoding, whereas the counterfactual branch recomputes the complete multimodal prefix. To verify that this execution-path difference is not itself responsible for LOCI's changes, we run the complete counterfactual code path with $\lambda = 0$. The hook remains registered and all image-token indices, attention masks, position IDs, model dtypes, and deterministic decoding processors are identical, but Eq. 1 becomes an exact no-op.

For prefix event e, let $\mathbf{z}_e^{\text{cache}} \in \mathbb{R}^{|\mathcal{V}|}$ be the raw finite-vocabulary logits from cached decoding and let $\mathbf{z}_e^{\text{full},0}$ be the logits from the full-forward no-op path. We report

$$\text{MeanAbsDiff} = \frac{\sum_e \sum_{v\in\mathcal{V}} \left| z_{e,v}^{\text{cache}} - z_{e,v}^{\text{full},0} \right|}{|\mathcal{E}_{\text{id}}|\,|\mathcal{V}|},$$

and

$$\text{MaxAbsDiff} = \max_{e,v} \left| z_{e,v}^{\text{cache}} - z_{e,v}^{\text{full},0} \right|.$$

Top-1 agreement is computed after applying the same repetition penalty, token masks, and stopping constraints to both score vectors. For the sequence-level check, we generate reports with LOCI at $\lambda = 0$ and compare the generated token-ID sequences with standard cached decoding. This test is run independently for all three backbones on a fixed validation subset and is not used for model selection.

Fixed event set for margin diagnostics. Candidate-level comparisons across true, shuffled, and blank images can be confounded if each condition produces a different set of triggers. We therefore construct a fixed event set separately on the validation and test splits. An event

$$e = (i, Y_{<t}, y, \phi)$$

contains a study identifier, a prefix obtained from standard greedy decoding with the true image, a positive-scope matched candidate, and its canonical finding. The same prefix, candidate, and finding are then evaluated using the true image, a fixed shuffled image from a different study, and a constant blank image. Thus, the three input conditions contain exactly the same labeled events.

For event e under condition κ, the base score and counterfactual margin are

$$b_e^{(\kappa)} = \log P_\theta\big(y_e \mid I_e^{(\kappa)}, Y_{<t_e}\big), \qquad m_e^{(\kappa)} = b_e^{(\kappa)} - \log P_\theta^{\mathrm{cf},\phi_e}\big(y_e \mid I_e^{(\kappa)}, Y_{<t_e}\big),$$

where $P_\theta^{\mathrm{cf},\phi_e}$ is the next-token distribution obtained by applying Eq. 1 with $\hat{u}_{\phi_e}$ at layer ℓ to the image tokens of $I_e^{(\kappa)}$. Both terms are log-probabilities of the complete next-token distribution over $\mathcal{V}$. Under $\kappa = \mathrm{true}$ these coincide exactly with $\log p_t^b(y_e)$ and $m_t(y_e, \phi_e)$ of Eq. 2; the other two conditions re-evaluate the same quantities with a substituted image.

Reference finding labels are mapped to $r_e \in \{0, 1\}$, where POS is positive and NEG, UNC, and BLANK are non-positive. Events whose findings cannot be mapped to the fixed label set are excluded from this AUC analysis but retained in trigger statistics.

Incremental predictive-value models. To control for finding-specific prevalence and score scale, base scores and margins are standardized within finding using validation-set statistics:

$$\tilde{b}_e = \frac{b_e - \mu_{b,\phi_e}}{\sigma_{b,\phi_e} + \epsilon}, \qquad \tilde{m}_e = \frac{m_e - \mu_{m,\phi_e}}{\sigma_{m,\phi_e} + \epsilon}.$$

For each input condition, we fit the following unpenalized logistic models on validation events:

$$\mathcal{M}_b: \quad \mathrm{logit}\, P(r_e = 1) = \beta_0 + \gamma_{\phi_e} + \beta_b \tilde{b}_e,$$

$$\mathcal{M}_m: \quad \mathrm{logit}\, P(r_e = 1) = \beta_0 + \gamma_{\phi_e} + \beta_m \tilde{m}_e,$$

and

$$\mathcal{M}_{b+m}: \quad \mathrm{logit}\, P(r_e = 1) = \beta_0 + \gamma_{\phi_e} + \beta_b \tilde{b}_e + \beta_m \tilde{m}_e,$$

where γ_{ϕ_e} is a finding-specific intercept. Coefficients are frozen before evaluating the corresponding test events.

The incremental statistic is

$$\Delta\mathrm{AUC} = \mathrm{AUC}(\mathcal{M}_{b+m}) - \mathrm{AUC}(\mathcal{M}_b).$$

Confidence intervals resample studies and retain every event belonging to each sampled study. The same bootstrap studies are used for $\mathcal{M}_b$ and $\mathcal{M}_{b+m}$ so that the interval is paired.

Table 28: Search grids and selected settings for LOCI and clinical controls on MIMIC-CXR with Qwen2.5-VL. All settings are selected on the validation split using CheXbert-5 F1, with RadGraph F1 used as a tie-breaker. The selected setting is fixed before test evaluation.

Method	Tuned parameters	Search grid	Selected setting
Tuned conservative prompt	Prompt variant	$\{P_0, P_1, P_2, P_3, P_4\}$	$P_* = P_3$
Scope-aware lexicon penalty	β	$\{0.05, 0.10, 0.20, 0.30, 0.50, 0.80, 1.00\}$	$\beta_* = 0.20$
Blank-image contrastive	$\alpha_{\mathrm{blank}}, \alpha_{\mathrm{spc}}$	$\alpha_{\mathrm{blank}} \in \{0.10, 0.20, \ldots, 1.00\}$; $\alpha_{\mathrm{spc}} \in \{0.30, 0.50, 0.70\}$	$(\alpha_{\mathrm{blank}}, \alpha_{\mathrm{spc}}) = (0.80, 0.30)$
Random-direction intervention	α, λ	$\alpha \in \{0.10, 0.20, \ldots, 1.00\}$; $\lambda \in \{0.50, 1.00, 1.50, 2.00\}$; $\alpha_{\mathrm{spc}} = 0.30$ (fixed)	$(\alpha, \lambda, \alpha_{\mathrm{spc}}) = (0.80, 1.00, 0.30)$
Scope-aware direct erasure	η	$\{0.25, 0.50, 1.00\}$	$\eta_* = 0.50$
LOCI	$\alpha, \lambda, \alpha_{\mathrm{spc}}, K_{\mathrm{cand}}, G_{\max}, \ell$	$\alpha \in \{0.10, 0.20, \ldots, 1.00\}$; $\lambda \in \{0.50, 1.00, 1.50, 2.00\}$; $\alpha_{\mathrm{spc}} \in \{0.30, 0.50, 0.70\}$; $K_{\mathrm{cand}} \in \{5, 10, 15, 20, 30, 40\}$; $G_{\max} \in \{1, 2, 3\}$; $\ell \in \{0, 4, 7, 11, 14, 18, 21, 25, 27\}$	$(0.80, 1.00, 0.30, 30, 3, 18)$

D ADDITIONAL ANALYSIS

This section reports supplementary analyses that support the main findings. These analyses are descriptive and are not used for prompt selection, hyperparameter selection, or model selection.

D.1 CLINICAL–LEXICAL TRADE-OFF

As a lexical-overlap diagnostic, Table 29 reports ROUGE-L changes relative to standard decoding. The ROUGE-L deltas are small and sometimes negative even when RadGraph and CheXbert improve. This supports the interpretation that LOCI changes clinically salient assertions rather than simply making generated reports more lexically similar to references.

D.2 Full Controlled Input Perturbation Results

Table 30 provides the full results corresponding to the input-perturbation deltas reported in the main text. The true-image condition shows the largest LOCI gains, while blank-image and shuffled-image conditions show much smaller improvements. This supports the interpretation that LOCI's counterfactual signal depends on image-conditioned hidden states, although the evidence remains automatic and proxy-based.

Table 29: Clinical–lexical trade-off of LOCI relative to standard decoding on chest X-ray report generation (Δ: absolute change in percentage points; higher is better).

		Clinical (Δ)		**Lexical** (Δ)
Backbone	**Data**	R-F1	5c-F1	ROUGE-L
LLaVA-Med	MIMIC	$+1.43$	$+1.93$	-0.44
	IU	$+1.68$	$+2.12$	$+0.12$
Hulu-Med	MIMIC	$+0.44$	$+0.68$	-0.56
	IU	$+2.74$	$+3.18$	-0.23
Qwen2.5-VL	MIMIC	$+1.22$	$+5.33$	$+0.38$
	IU	$+0.27$	$+0.30$	$+0.42$

Table 30: Full controlled input perturbation results on the MIMIC-CXR single-view subset with Qwen2.5-VL. The reference report is kept fixed while the input image is either the true radiograph, a constant blank image, or a shuffled radiograph from another test study. Shuffled images are sampled from different studies. When MIMIC-CXR view-position metadata are available for the source example, shuffling is performed within the same view-position group; otherwise, the shuffled image is sampled from all different-study test images.

Input	Method	R-F1	14c-F1	5c-F1	Pos-Prec	Pos-Rec	Len	POS/report	UNC/report	NEG/report
True image	Standard	9.70	11.15	17.20	15.02	20.20	46.2	0.83	0.07	2.41
	LOCI	11.05	13.55	23.00	23.44	22.71	45.5	0.60	0.08	2.55
Blank image	Standard	8.02	9.45	13.24	12.8	18.6	44.0	0.82	0.07	2.38
	LOCI	8.27	9.95	14.09	14.3	18.0	43.5	0.76	0.08	2.42
Shuffled image	Standard	8.14	9.65	13.55	12.9	18.8	44.8	0.83	0.08	2.40
	LOCI	8.59	10.55	15.00	15.3	18.7	44.2	0.77	0.09	2.44

D.3 Additional per-category results

Tables 31 and 32 provide per-category CheXbert analyses on MIMIC-CXR with Qwen2.5-VL. The largest gains appear on clinically frequent positive findings such as pleural effusion, consolidation, edema, and atelectasis. The positive-label frequency does not collapse uniformly, which is consistent with the main-text observation that LOCI is not simply suppressing all positive findings. These category-wise analyses are intended to characterize where LOCI changes CheXbert agreement. They should not be interpreted as radiologist-adjudicated evidence of correctness for individual findings.

D.4 Triggered findings and category-wise counterfactual margins

Table 33 summarizes the most frequent positive-scope LOCI triggers on MIMIC-CXR with Qwen2.5-VL. Table 34 reports category-wise counterfactual margin diagnostics for triggered candidates whose canonical findings can be mapped to CheXbert observations. The margin is more discriminative for categories such as pleural effusion and consolidation, and weaker for broad categories such as lung opacity and lung lesion. This pattern is expected because broad opacity-like categories are visually and linguistically heterogeneous.

Table 34 also suggests that the scale of the counterfactual signal is consistent and not dominated by geometric artifacts of specific concept directions, ensuring that reranking is driven by instance-specific evidence rather than concept-level bias.

Table 31: Per-category CheXbert-5 analysis on MIMIC-CXR with Qwen2.5-VL. Cat-F1 is the category-level CheXbert F1. Pos-Prec and Pos-Rec treat POS as positive and merge NEG, UNC, and BLANK as non-positive. Deltas are LOCI minus standard decoding.

Category	Std Cat-F1	LOCI Cat-F1	ΔF1	Std Pos-Prec	LOCI Pos-Prec	Std Pos-Rec	LOCI Pos-Rec
Atelectasis	16.80	22.64	+5.84	13.4	21.82	22.5	23.52
Cardiomegaly	23.70	25.77	+2.07	29.5	34.37	19.8	20.61
Consolidation	13.00	19.04	+6.04	10.5	18.32	17.0	19.82
Edema	14.60	19.92	+5.32	11.3	18.14	20.5	22.07
Pleural Effusion	19.65	27.02	+7.37	17.0	28.14	23.2	25.99

Table 32: Per-category CheXbert-14 deltas on MIMIC-CXR with Qwen2.5-VL. All values are LOCI minus standard decoding. Cat-F1 is the category-level CheXbert F1. Pos-Prec and Pos-Rec treat POS as positive and merge NEG, UNC, and BLANK as non-positive.

Category	ΔCat-F1	ΔPos-Prec	ΔPos-Rec
Enlarged Cardiomediastinum	+0.37	+1.0	-0.5
Cardiomegaly	+2.07	+4.87	+0.81
Lung Opacity	+1.54	+2.4	+3.0
Lung Lesion	+0.43	+1.2	+0.5
Edema	+5.32	+6.84	+1.57
Consolidation	+6.04	+7.82	+2.82
Pneumonia	+0.12	+0.5	-0.8
Atelectasis	+5.84	+8.42	+1.02
Pneumothorax	+0.63	+3.5	+0.7
Pleural Effusion	+7.37	+11.14	+2.79
Pleural Other	+0.06	+0.2	-0.2
Fracture	+0.00	+0.0	+0.0
Support Devices	+0.69	+1.0	+1.5
No Finding	+0.32	+0.5	+0.6

D.5 Error-side trigger coverage

The trigger statistics in Table 7 measure how often LOCI is invoked during decoding, but do not quantify what fraction of clinically relevant baseline errors are reachable by the fixed phrase lexicon and scope gate. We therefore perform a post-hoc error-side coverage analysis on PadChest-GR using the RadFact logical-entailment judgments summarized in Appendix A.1.

For each standard-decoding report, the RadFact-based evaluator identifies generated positive findings that are either matched to a radiologist- annotated reference finding or counted as unsupported positives. We retain the generated surface span associated with each extracted positive finding and replay the standard-decoding token sequence through the frozen LOCI phrase matcher and scope gate.

Let $\mathcal{F}_i^{\mathrm{FP}}$ denote the set of unsupported positive findings produced by standard decoding for image i, and let $\mathcal{F}_i^{\mathrm{TP}}$ denote the standard-decoding positive findings matched to reference annotations. For a generated finding mention f, define

$$T(f) = \mathbf{1}[f \text{ contains a phrase completion assigned POS by the LOCI trigger}].$$

We define unsupported-positive trigger coverage as

$$\text{FP-Coverage} = \frac{\sum_i \sum_{f \in \mathcal{F}_i^{\mathrm{FP}}} T(f)}{\sum_i |\mathcal{F}_i^{\mathrm{FP}}|}.$$

Analogously, grounded-positive exposure is

$$\text{TP-Exposure} = \frac{\sum_i \sum_{f \in \mathcal{F}_i^{\mathrm{TP}}} T(f)}{\sum_i |\mathcal{F}_i^{\mathrm{TP}}|}.$$

Table 33: Most frequent LOCI-triggered findings on MIMIC-CXR with Qwen2.5-VL. Trigger share is computed over POS-triggered finding groups. Changed-step share is the fraction of triggers for that finding where LOCI changes the standard greedy decision. Selected-after-rerank is the fraction of triggers where a matched candidate for the finding is selected after reranking.

Finding	Trigger share (%)	Changed-step share (%)	Selected-after-rerank (%)
Pleural Effusion	20.8	27.8	56.2
Atelectasis	15.6	20.1	47.5
Consolidation	12.7	32.4	58.7
Edema	10.2	24.6	51.4
Cardiomegaly	8.9	17.2	44.8
Pneumothorax	5.1	23.5	49.1
Lung Opacity	16.9	18.7	46.0
Lung Lesion	4.2	29.0	41.3

Table 34: Category-wise counterfactual margin diagnostic on MIMIC-CXR with Qwen2.5-VL. We report triggered candidates whose canonical finding maps to a CheXbert category.

Category	Positive events	Non-positive events	Margin positive	Margin non-positive	AUC
Atelectasis	640	1410	0.34	0.14	0.64
Cardiomegaly	480	800	0.32	0.13	0.63
Consolidation	310	1300	0.48	0.16	0.69
Edema	360	850	0.41	0.14	0.67
Pleural Effusion	710	1720	0.53	0.15	0.72
Pneumothorax	85	605	0.40	0.12	0.66
Lung Opacity	930	1460	0.27	0.17	0.59
Lung Lesion	120	440	0.25	0.18	0.56

To measure the outcome conditional on an intervention opportunity, we intersect these sets with the grounded transition analysis of Table 9. Correction@covered is

$$\text{Correction@covered} = \frac{\#\{\text{covered standard unsupported positives corrected by LOCI}\}}{\#\{\text{covered standard unsupported positives}\}},$$

and Suppression@exposed is

$$\text{Suppression@exposed} = \frac{\#\{\text{exposed standard positives suppressed by LOCI}\}}{\#\{\text{exposed standard positives}\}}.$$

All grounded labels, ontology mappings, and one-to-one matching rules are identical to the primary PadChest-GR evaluation. This analysis is performed only after generation and is not used for lexicon construction, hyperparameter selection, or model selection.

D.6 Single-view versus multi-view subset analysis

To quantify the impact of the image-level evaluation protocol on multi-view studies, we perform a subset analysis on MIMIC-CXR with Qwen2.5-VL. We partition the study-level test split before image-level expansion. A study is labeled *single-view* if exactly one radiograph is associated with the report, and *multi-view* otherwise.

For the single-view subset, each study contributes exactly one image–report pair, so the cross-view input–reference mismatch discussed in Section 4.1 is absent. For the multi-view subset, we keep the main-paper protocol and evaluate each image independently against the corresponding study report. All prompts, preprocessing, decoding settings, and evaluator configurations are identical to those used in the main experiments.

Table 35 reports results for standard decoding and LOCI on the two subsets. LOCI retains gains on the single-view subset, indicating that the improvements are not explained solely by the multi-view evaluation artifact. At the same time, results on the multi-view subset should still be interpreted cautiously as reference-based agreement under a single-image protocol rather than as definitive evidence of per-image grounding.

Table 35: Single-view versus multi-view subset analysis on MIMIC-CXR with Qwen2.5-VL. We partition the study-level test split before image-level expansion. Single-view studies contain exactly one image, so $\#\mathrm{Studies} = \#\mathrm{Pairs}$ and the cross-view input–reference mismatch is absent. Multi-view studies are then evaluated with the main-paper protocol, i.e., each image is processed independently against the corresponding study report.

Subset	Method	#Studies	#Pairs	R-F1	14c-F1	5c-F1
Single-view	Standard	1742	1742	9.70	11.15	17.20
	LOCI	1742	1742	11.05	13.55	23.00
Multi-view	Standard	1527	3417	9.93	11.58	18.04
	LOCI	1527	3417	11.10	13.51	22.45
Δ (LOCI $-$ Standard), single-view		–	–	+1.35	+2.40	+5.80
Δ (LOCI $-$ Standard), multi-view		–	–	+1.17	+1.93	+4.41