HIDDEN IN THE DECODES: RECOVERING CLINICAL FINDINGS WITHOUT RETRAINING MEDICAL MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic radiology report generation commonly commits to a single decoded report from a medical multimodal large language model (MLLM). However, clinically correct findings omitted from this report may be hidden in other stochastic decodes produced by the same frozen model, while unsupported findings may also recur across decodes. Existing approaches address these errors by modifying training objectives, model architectures, supervision, or refinement modules, but typically require model updates or additional training and do not ensure that a single decode captures all supported findings. We introduce HIDE, a model-agnostic, training-free framework for recovering clinical findings without retraining medical MLLMs. HIDE treats the frozen generator as a proposal distribution for verification-guided inference: stochastic decoding broadens candidate coverage, cross-sample consistency provides model-side proposal support, localized visual support from a prototype bank verifies image compatibility, and source-constrained minimal edits compose only jointly supported findings. Given an initial draft and additional stochastic samples, HIDE extracts finding-level candidates and selects among three conservative actions: adding a supported finding missing from the draft, removing an unsupported draft claim, or abstaining when the evidence is inconclusive. Across six frozen MLLM report generators, HIDE achieves mean relative gains of 27.3%, 54.1%, and 16.0% on CheXbert-F1-14, CheXbert-F1-5, and RadGraph-F1 on MIMIC-CXR, respectively. On CheXpert Plus, the corresponding gains reach 41.2%, 239.7%, and 19.4%. The correction signal also transfers to IU-Xray when fitted only on MIMIC-CXR or CheXpert Plus. HIDE further achieves state-of-the-art performance among training-free correction methods.

1 INTRODUCTION

Automated radiology report generation aims to reduce radiologists' burden of writing clinical reports from medical images (Hyland et al., 2023). With paired image–report supervision, multimodal large language models (MLLMs) learn a conditional distribution $p_\theta(y \mid x)$ over report sequences y conditioned on an input image x. Because image–report supervision can be incomplete, noisy, or weakly grounded in visual evidence, this learned distribution may not accurately reflect the distribution of image-supported clinical reports. Existing generators therefore produce reports by decoding from an imperfect learned distribution rather than directly from the image-grounded clinical distribution. At inference, autoregressive decoding selects or samples a single token sequence from this distribution (Bannur et al., 2024). The resulting report may contain hallucinated findings or omit clinically relevant abnormalities (Goswami et al., 2025). Hallucinated findings can arise when the learned distribution is insufficiently grounded in the image and assigns probability to clinically plausible but unsupported findings (Chen et al., 2026a). Conversely, an image-supported finding may have non-zero probability under the learned distribution but be absent from a particular decoded report when sequences expressing that finding receive lower probability than competing sequences (Holtzman et al., 2019). These two error modes remain a longstanding challenge in automated report generation. Prior studies have reported that hallucinated findings and omitted abnormalities can compromise the clinical utility of generated reports (Goswami et al., 2025).

To address these errors, existing approaches improve report generation through three complementary strategies: training-objective methods, grounding-based methods, and refinement or self-critique

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

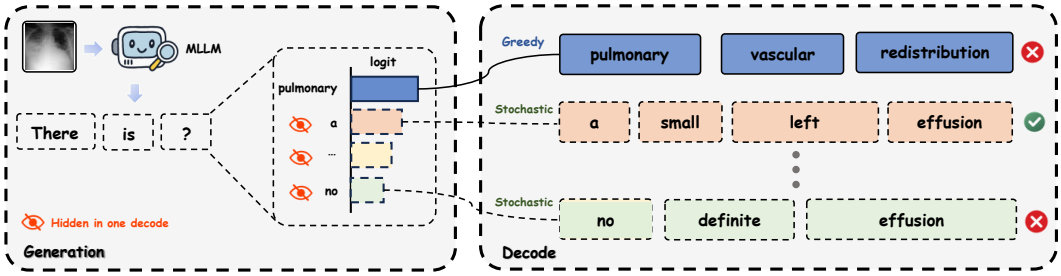

Figure 1: Greedy and stochastic decoding can produce different candidates from the same image.

methods. These interventions aim to make the learned report distribution more complete, consistent, and better aligned with image evidence. First, training-objective methods modify the learning signal through clinically informed rewards, contrastive objectives, or preference optimization to encourage more complete, consistent, or preference-aligned reports (Miura et al., 2021; Delbrouck et al., 2022; Xiao et al., 2025; Liu et al., 2025). Second, grounding-based methods improve image–report correspondence by incorporating anatomical regions, cross-modal alignment, grounded-report supervision, or segmentation-derived visual information (Tanida et al., 2023; Bannur et al., 2024; Huang et al., 2025; Sharma et al., 2025). Third, refinement and self-critique methods re-evaluate generated content against image representations or localized abnormality hypotheses (Kapadnis et al., 2024; Yan et al., 2026). Yet hallucinations can persist in individual decodes, and the model updates or training used by these methods do not ensure that a single report captures every correct finding.

However, improving the generator does not guarantee that a single decoded report captures all correct findings it can produce. At each decoding step, the MLLM assigns probability mass to multiple next-token continuations: greedy decoding follows a single high-probability path, whereas stochastic decoding samples alternative paths. A correct finding with non-zero probability may therefore be absent from one decode, while repeated stochastic decoding can expose both correct and incorrect candidates (Figure 1). We observe this behavior empirically: correct findings omitted from an initial draft can appear in other reports sampled from the same frozen generator for the same image, with different reports contributing complementary findings. An omission in the initial draft therefore does not necessarily mean that the generator cannot produce the missing finding. These additional stochastic samples provide candidate findings for correcting omissions without retraining the generator. However, unsupported findings may also recur across samples when they are favored by the learned distribution. The challenge then becomes: *How can we determine which candidate findings are supported by the image and should be incorporated into the report?*

To address this challenge, we introduce HIDE, a model-agnostic, training-free framework for correcting chest X-ray reports without retraining the generator. Given an initial draft and additional stochastic samples from the same frozen MLLM, HIDE extracts candidate findings, including those absent from the initial draft, and associates them with relevant anatomical locations. Stochastic decoding broadens candidate coverage, while cross-sample consistency measures how consistently each candidate finding is proposed across samples and provides model-side proposal support. A localized prototype bank provides localized visual support at the associated anatomical location. Indexed by finding, anatomical location, and imaging projection, the prototype bank contains positive, normal, and confounder references for local visual comparison. A consensus predictor combines cross-sample consistency and localized visual support as a finding-level verifier and selects among three conservative actions: adding a supported finding missing from the initial draft, removing an unsupported draft claim, or abstaining when the evidence is inconclusive. The selected actions are applied as source-constrained minimal edits, with removals restricted to claims in the original draft.

Our contributions are summarized as follows:

- We prove that stochastic decoding from a frozen medical MLLM can expose complementary clinical findings omitted by an initial greedy report, establishing a training-free correction setting with respect to the report generator.
- We introduce HIDE, which combines finding-level cross-sample construction, a localized prototype bank with positive, normal, and confounder references, an evidence-guided commonality predictor, and source-constrained minimal editing to add supported findings and remove unsupported claims.

- We evaluate HIDE on three datasets and six frozen medical MLLM generators. On MIMIC-CXR, HIDE achieves mean relative gains of 27.3%, 54.1%, and 16.0% on CheXbert-F1-14, CheXbert-F1-5, and RadGraph-F1, respectively; on CheXpert Plus, the corresponding gains reach 41.2%, 239.7%, and 19.4%. The correction signal also transfers to IU-Xray without target-domain fitting or generator updates.

2 RELATED WORK

MLLMs for Medical Report Generation. Recent medical multimodal large language models (MLLMs) have progressed from CXR-specific findings generation toward broader medical understanding and reasoning. LLaVA-Rad (Zambrano Chaves et al., 2025) demonstrates that an open-source small multimodal model can generate free-text findings from chest X-rays, offering a more accessible alternative to much larger medical foundation models. CheXagent (Chen et al., 2024) extends this CXR-focused line of work to multiple interpretation tasks and evaluates report drafting in radiologist workflows. MedGemma (Sellergren et al., 2025) further extends medical vision–language modeling through an open collection of medically tuned models; its technical report evaluates chest X-ray report generation alongside image classification, visual question answering, and agentic tasks. Lingshu (Xu et al., 2026) targets remaining gaps in medical knowledge coverage, hallucination susceptibility, and complex reasoning, and evaluates a generalist medical MLLM on multimodal question answering, textual question answering, and radiology report generation. These models alone cannot guarantee that generated reports are complete and grounded in the image, motivating subsequent work on factuality and correction.

Factuality and Correction in Radiology Report Generation. Recent methods target factuality at different stages of report generation. DART (Park et al., 2025) uses disease-aware image-to-text retrieval followed by self-correction to re-align the initial report with the input X-ray. LLaVA-TA (Cheng & Subramanian, 2026) reduces reliance on narrative priors by generating independent topic-level findings conditioned on the full image and corresponding anatomy. RadSCR (Yan et al., 2026) verifies localized abnormality proposals against alternative abnormalities, patient X-rays, and potential false negatives before report assembly. OraPO (Chen et al., 2026b) targets factuality during training by converting failed GRPO explorations into preference supervision and rewarding atomic clinical facts entailed by ground-truth labels. FactCheXcker (Heiman et al., 2025) targets measurement hallucinations with a modular query–code–update pipeline. These methods seek to shift the generator toward more factually grounded outputs, but such shifts do not ensure that a single decoded report captures every image-supported finding the model can produce. Our work instead recovers omitted findings from additional stochastic outputs of the same frozen generator and verifies them using cross-sample and localized visual evidence.

Cross-Sample Consistency and Selection. A fixed generator may produce complementary reports across stochastic runs, motivating inference-time use of multiple samples. Best-of-N inference (Chow et al., 2025) generates multiple responses and selects one with a verifier, while CCS (Zhang et al., 2026a) selects the candidate with the highest text–image clinical consensus. SCUQ-RRG (Wang et al., 2025b) uses semantic consistency across samples to estimate report- and sentence-level uncertainty, and RadFlag (Zhang et al., 2025) flags claims lacking cross-sample support as potential hallucinations. In contrast, CCD (Zhang et al., 2026b) and related contrastive-decoding methods alter token logits or internal representations during a single decoding path. These methods operate at the report or decoding-path level and do not explicitly recover complementary findings across samples. Our method combines cross-sample recurrence with localized visual evidence to recover image-supported findings omitted from the initial draft.

3 METHOD

We present HIDE, a framework for correcting radiology reports without updating the report generator. Given an image, HIDE extracts finding-level candidates from an initial draft and additional stochastic generations, combines cross-sample commonality with localized visual evidence, and applies source-constrained minimal edits to the draft.

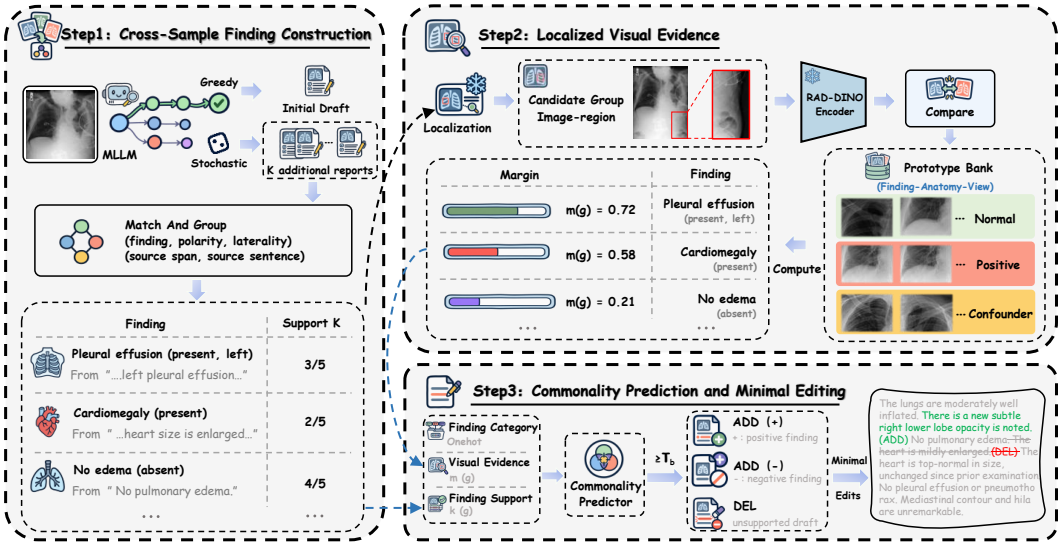

Figure 2: **Overview of HIDE.** Candidates are extracted from stochastic decodes, evaluated by the commonality predictor and localized visual evidence, incorporated through source-constrained edits.

3.1 Problem Formulation

For an image x, a frozen report generator G_θ produces an initial draft $y^{(0)}$ using greedy decoding and K additional reports $\{y^{(k)}\}_{k=1}^{K}$ using stochastic decoding. Let $Y \sim q_\theta(\cdot \mid x)$ denote a report sampled under the chosen stochastic decoding parameters, and define $p_f = \Pr(f \in \mathrm{Find}(Y) \mid x)$ as the probability that a stochastic sample expresses the finding f. Because greedy decoding produces one deterministic decoding outcome, the absence of f from the initial draft $y^{(0)}$ does not imply that $p_f = 0$. When $p_f > 0$, each stochastic sample independently expresses f with probability p_f. Under independent draws, a sample omits f with probability $1 - p_f$, so the probability that all K samples omit f is $(1 - p_f)^K$.Taking the complement, the probability that f appears in at least one of the K additional reports is:

$$P_K(f) = 1 - (1 - p_f)^K. \tag{1}$$

For $0 < p_f < 1$, this probability increases monotonically with K: repeated sampling does not create a finding that has zero probability under the sampling distribution, but makes an already generatable finding more likely to become observable. Temperature and top-p affect this distribution, and therefore p_f, whereas K controls the number of independent opportunities to observe f. We use the additional reports to construct candidate findings and then use image evidence to determine which candidates should modify the initial draft.

3.2 Cross-Sample Finding Construction

We extract structured findings from the initial draft and additional reports using a fixed extraction procedure. The model and prompt are specified in Appendix C. Each finding is represented by:

$$e = (f, p, \ell, c, s), \tag{2}$$

where f is the finding category, $p \in \{\text{present}, \text{absent}, \text{uncertain}\}$ denotes its polarity, ℓ is the stated laterality, c is the source span, and s is the source sentence. The tuple (f, p, ℓ) matches equivalent findings across reports, while (c, s) preserves text required for source-constrained editing. Anatomical regions are localized using phrase grounding on extracted finding spans.

Let $E^{(k)}$ denote the set of extracted findings from report k. We measure cross-sample support for a finding unit by counting the number of stochastic reports in which it appears:

$$n(f, p, \ell) = \sum_{k=1}^{K} \mathbf{1}\left[(f, p, \ell) \in E^{(k)}\right], \tag{3}$$

where $\mathbf{1}[\cdot]$ is the indicator function, equal to 1 when its argument is true and 0 otherwise. Each report contributes at most once, and findings with identical (f, p, ℓ) are grouped while retaining their source

spans. We form three correction branches: ADD^+ contains present or uncertain findings expressed in additional reports but absent from the draft; ADD^- contains explicitly absent findings not covered by the draft; and DEL contains present or uncertain findings already expressed in the draft.

3.3 Localized Prototype-Based Visual Evidence

Cross-sample support measures consistency across generations, but it cannot establish whether a finding is present in the image. To bridge this gap, we construct a finding-specific prototype bank from the training split of Chest ImaGenome, using its scene-graph labels and anatomical bounding boxes. Each bank entry is indexed by finding category, anatomical region, and projection, and is assigned one of three roles: positive, normal, or confounder. Positive prototypes contain the target finding, normal prototypes lack it, and confounder prototypes contain visually similar abnormalities without the target finding.

For a candidate group g, we crop the corresponding image region and use the frozen RAD-DINO encoder to obtain a normalized embedding. We retrieve the J nearest prototypes for each role using cosine similarity and let $S^t(g)$ denote the mean similarity for role $t \in \{\mathrm{pos}, \mathrm{norm}, \mathrm{conf}\}$. We summarize relative visual support with the prototype margin:

$$m(g) = S^{\mathrm{pos}}(g) - \frac{1}{\beta} \log\big(\exp(\beta S^{\mathrm{norm}}(g)) + \exp\big(\beta S^{\mathrm{conf}}(g)\big)\big) . \tag{4}$$

The margin is used as an image-evidence feature for candidate prediction, rather than interpreted as a calibrated disease probability.

3.4 Commonality Prediction and Candidate Authorization

The candidate pool contains findings with different meanings for different correction branches, so a fixed support threshold cannot treat all candidates identically. HIDE therefore introduces a branch-specific commonality predictor that learns how cross-sample recurrence should be interpreted together with the finding category and localized image evidence. For each branch $b \in \{\mathrm{ADD}^+, \mathrm{ADD}^-, \mathrm{DEL}\}$, we form:

$$\phi_b(g) = [\mathrm{onehot}(f_g),\, k_b(g),\, m(g)] , \qquad q_b(g) = T_b(\phi_b(g)) , \tag{5}$$

where $k_b(g)$ is the branch-specific commonality feature derived from the support count $n(g)$, and T_b is the predictor for branch b. The ADD^+ predictor estimates whether a candidate is supported by the image, whereas the ADD^- and DEL predictors estimate whether the corresponding claim lacks image support. Separate predictors are used because the three branches have different action objectives. A candidate is authorized when:

$$a_b(g) = \mathbf{1}[q_b(g) \geq \tau_b] , \tag{6}$$

where τ_b is a branch-specific threshold. At the finding level, opposite-polarity additions are resolved by their scores and contradictory edits are discarded when required dependencies are not satisfied.

3.5 Source-Constrained Minimal Editing

The correction decisions are applied through an editor that revises $y^{(0)}$ rather than regenerating a report. For an authorized addition, the editor reuses the source sentence from an additional report corresponding to the selected finding. For an authorized deletion, it removes only the matched finding span from the initial draft. All edits are constrained by (c, s), preserving unaffected text while keeping additions tied to source sentences and deletions tied to matched spans.

4 Experiments

Baseline. We evaluate HIDE on six publicly released medical MLLM checkpoints: MedR1-2B (Lai et al., 2026), MedVLM-R1-2B (Pan et al., 2025), MedGemma-4B-IT (Sellergren et al., 2025), Lingshu-7B (Xu et al., 2026), LLaVA-Rad-7B (Zambrano Chaves et al., 2025), and CheXagent-2-3B (Chen et al., 2024). We compare HIDE with greedy decoding and five training-free baselines: DeCo (Wang et al., 2025a), ICD (Wang et al., 2024), VCD (Leng et al., 2024), CCD (Zhang et al., 2026b), and CCS (Zhang et al., 2026a).

Table 1: **Results on MIMIC-CXR and CheXpert Plus.** Relative gains of HIDE over the corresponding base model are shown in parentheses. **Bold** denotes the better result within each base-HIDE pair. The final row of each dataset reports the mean relative gain across the six report generators.

Model	Radiology-specific Metrics					General NLG Metrics					
	CheXbert_{14} ↑	CheXbert_{5} ↑	RadGraph-F1 ↑	RaTEScore ↑	RadEval-BERT ↑	ROUGE-L ↑	B-1 ↑	B-4 ↑	METEOR ↑	CIDEr ↑	BERTScore ↑
MIMIC-CXR											
Lingshu-7B	0.3439	0.3708	0.2021	0.5122	0.2075	0.2747	0.1737	0.0232	0.1683	0.9822	0.5037
+ HIDE	**0.4591** (+33.5%)	**0.4863** (+31.1%)	**0.2196** (+8.7%)	**0.5261** (+2.7%)	**0.2131** (+2.7%)	**0.2909** (+5.9%)	**0.2240** (+29.0%)	**0.0256** (+10.3%)	**0.1947** (+15.7%)	**1.0122** (+3.1%)	**0.5137** (+2.0%)
MedGemma-4B-IT	0.4783	0.4584	0.2008	0.5252	0.2090	0.2555	0.2288	0.0279	0.2030	0.8107	0.4936
+ HIDE	**0.5143** (+7.5%)	**0.5173** (+12.8%)	**0.2269** (+13.0%)	**0.5513** (+5.0%)	**0.2097** (+0.3%)	**0.2828** (+10.7%)	**0.2813** (+22.9%)	**0.0320** (+14.7%)	**0.2299** (+13.3%)	**0.9076** (+12.0%)	**0.5092** (+3.2%)
MedVLM-R1-2B	0.2170	0.1386	0.1044	0.4419	0.1930	0.2148	0.2214	**0.0057**	0.1835	0.4164	0.3600
+ HIDE	**0.3165** (+45.9%)	**0.3165** (+128.4%)	**0.1193** (+14.3%)	**0.4618** (+4.5%)	**0.1974** (+2.3%)	**0.2206** (+2.7%)	**0.2220** (+0.3%)	**0.0057** (0.0%)	**0.1918** (+4.5%)	**0.4805** (+15.39%)	**0.3717** (+3.3%)
LLaVA-RAD-7B	0.5634	0.5741	0.2747	0.5494	0.2543	0.3248	0.2913	0.0643	0.2598	1.2695	0.5412
+ HIDE	**0.5858** (+4.0%)	**0.5977** (+4.1%)	**0.2798** (+1.9%)	**0.5545** (+0.9%)	**0.2576** (+1.3%)	**0.3303** (+1.7%)	**0.3036** (+4.2%)	**0.0655** (+1.9%)	**0.2647** (+1.9%)	**1.2855** (+1.3%)	**0.5439** (+0.5%)
Med-R1-2B	0.2098	0.1357	0.0883	0.4383	0.1899	0.2057	**0.2267**	0.0037	0.1740	0.3808	0.3455
+ HIDE	**0.3053** (+45.5%)	**0.2965** (+118.5%)	**0.1100** (+24.6%)	**0.4541** (+3.6%)	**0.1931** (+1.7%)	**0.2152** (+4.6%)	0.2258 (−0.4%)	**0.0040** (+8.1%)	**0.1841** (+5.8%)	**0.4255** (+11.7%)	**0.3616** (+4.7%)
CheXAgent-2-3B	0.4104	0.4411	0.1289	0.4278	0.1639	0.1521	0.0507	0.0052	0.0784	0.4425	0.3180
+ HIDE	**0.5220** (+27.2%)	**0.5731** (+29.9%)	**0.1722** (+33.6%)	**0.4678** (+9.4%)	**0.1709** (+4.3%)	**0.1934** (+27.2%)	**0.1240** (+144.6%)	**0.0094** (+80.8%)	**0.1208** (+54.1%)	**0.4930** (+11.4%)	**0.3457** (+8.7%)
Mean Rel. Gain	**+27.3%**	**+54.1%**	**+16.0%**	**+4.3%**	**+2.1%**	**+8.8%**	**+33.4%**	**+19.3%**	**+15.9%**	**+9.1%**	**+3.7%**
CheXpert Plus											
Lingshu-7B	0.2807	0.2970	0.1506	0.4653	0.1979	0.2210	0.1272	0.0103	0.1279	0.7440	0.4438
+ HIDE	**0.3769** (+34.3%)	**0.4228** (+42.4%)	**0.1815** (+20.5%)	**0.4742** (+1.9%)	**0.2000** (+1.1%)	**0.2352** (+6.4%)	**0.1574** (+23.7%)	**0.0109** (+5.8%)	**0.1434** (+12.1%)	**0.7931** (+6.6%)	**0.4569** (+3.0%)
MedGemma-4B-IT	0.4372	0.3368	0.1845	0.5070	0.2136	0.2391	0.2097	0.0180	0.1773	0.7665	0.4510
+ HIDE	**0.4867** (+11.3%)	**0.3810** (+13.1%)	**0.1998** (+8.3%)	**0.5143** (+1.4%)	**0.2193** (+2.7%)	**0.2478** (+3.6%)	**0.2246** (+7.1%)	**0.0181** (+0.6%)	**0.1916** (+8.1%)	**0.8221** (+7.3%)	**0.4535** (+0.6%)
MedVLM-R1-2B	0.1973	0.1379	0.0990	0.4371	0.2020	0.1979	0.1682	0.0023	0.1501	0.4038	0.3309
+ HIDE	**0.3083** (+56.3%)	**0.3115** (+125.9%)	**0.1217** (+22.9%)	**0.4581** (+4.8%)	**0.2143** (+6.1%)	**0.2023** (+2.2%)	**0.1774** (+5.5%)	**0.0030** (+30.4%)	**0.1631** (+8.7%)	**0.4683** (+16.0%)	**0.3405** (+2.9%)
LLaVA-RAD-7B	0.5113	0.5161	0.2230	0.5212	0.2251	0.2662	0.1976	0.0185	0.1740	0.9817	0.4806
+ HIDE	**0.5822** (+13.9%)	**0.5185** (+0.5%)	**0.2396** (+7.4%)	**0.5240** (+0.5%)	**0.2331** (+3.6%)	**0.2765** (+3.9%)	**0.2071** (+4.8%)	**0.0204** (+10.3%)	**0.1780** (+2.3%)	**1.0008** (+1.9%)	**0.4884** (+1.6%)
Med-R1-2B	0.1468	0.0250	0.0839	0.4269	0.2104	0.1911	0.1901	0.0015	0.1518	0.3791	0.3423
+ HIDE	**0.3175** (+116.3%)	**0.3333** (+1233.2%)	**0.1155** (+37.7%)	**0.4525** (+6.0%)	**0.2177** (+3.5%)	**0.2056** (+7.6%)	**0.1911** (+0.5%)	**0.0024** (+60.0%)	**0.1655** (+9.0%)	**0.4502** (+18.8%)	**0.3656** (+6.8%)
CheXAgent-2-3B	0.4851	0.4531	0.2089	0.4855	0.2094	0.1931	0.1014	0.0078	0.1060	0.6353	0.3222
+ HIDE	**0.5597** (+15.4%)	**0.5578** (+23.1%)	**0.2495** (+19.4%)	**0.5262** (+8.4%)	**0.2156** (+3.0%)	**0.2380** (+23.3%)	**0.1532** (+51.1%)	**0.0100** (+28.2%)	**0.1469** (+38.6%)	**0.6903** (+8.7%)	**0.3661** (+13.6%)
Mean Rel. Gain	**+41.2%**	**+239.7%**	**+19.4%**	**+3.8%**	**+3.3%**	**+7.8%**	**+15.5%**	**+22.5%**	**+13.1%**	**+9.9%**	**+4.7%**

Datasets and preprocessing. MIMIC-CXR (Johnson et al., 2019) and CheXpert Plus (Chambon et al., 2024) are used for corrector fitting and in-domain evaluation, with only 0.5% of the data used to fit correction components while keeping report generators frozen. We evaluate on the official MIMIC-CXR test split, CheXpert Plus validation split, and IU-Xray (Demner-Fushman et al., 2015) for cross-dataset transfer. Dataset preprocessing follows prior work (Wu et al., 2026; Zambrano Chaves et al., 2025). The prototype bank is built from Chest ImaGenome (Wu et al., 2021).

Evaluation metrics. We compare each generated *Findings* section with its reference using complementary text-based and clinical metrics. Text-based metrics include BLEU-1 and BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Denkowski & Lavie, 2014), CIDEr (Vedantam et al., 2015), and BERTScore (Zhang et al., 2019). These metrics assess lexical overlap, consensus-based similarity, and contextual semantic similarity. Clinical metrics include CheXbert-F1-14 and CheXbert-F1-5 (Smit et al., 2020), which measure F1 agreement over CheXbert-extracted observations using 14 and 5 classes, respectively; RadGraph-F1 (Yu et al., 2023), which evaluates overlap in clinical entities and relations; RaTEScore (Zhao et al., 2024), an entity-aware metric sensitive to medical synonyms and negation; and RadEval-BERT (Xu et al., 2025), a radiology-adapted semantic similarity metric. Higher scores indicate better agreement with the reference report.

Implementation details. All inference runs use four NVIDIA A800 GPUs. Candidate regions are localized with phrase grounding from MAIRA-2 (Bannur et al., 2024), and each bounding box is expanded about its center by approximately $1.2\times$ before cropping. Visual evidence is computed with the frozen RAD-DINO encoder (Pérez-García et al., 2025). The model-specific parameters and sampling budget K are reported in Appendix A, and the prototype filtering procedure is described in Appendix B. The finding extraction configuration is detailed in Appendix C. The branch-specific thresholds for commonality prediction, τ_{ADD^+}, τ_{ADD^-}, and τ_{DEL}, are reported in Appendix D.

4.1 Main Results

We show that HIDE improves reports generated by frozen models through stochastic generations and localized visual verification on MIMIC-CXR and CheXpert Plus. As shown in Table 1, HIDE produces positive mean relative gains across all 11 metrics and six generators on both datasets. Largest improvements occur on clinical metrics: on MIMIC-CXR, mean gains are 27.3% for CheXbert-F1-14 and 16.0% for RadGraph-F1, while on CheXpert Plus they reach 41.2% and 19.4%, respectively. CheXbert-F1-5 increases by 54.1% on MIMIC-CXR and 239.7% on CheXpert Plus; the latter is partly influenced by the low baseline score of Med-R1. Reference-based metrics also improve, with gains of 9.1% and 9.9% in CIDEr and 8.8% and 7.8% in ROUGE-L on MIMIC-CXR and CheXpert Plus, respectively. These results indicate that HIDE can recover useful findings from stochastic generations and improve the initial report without retraining the generator.

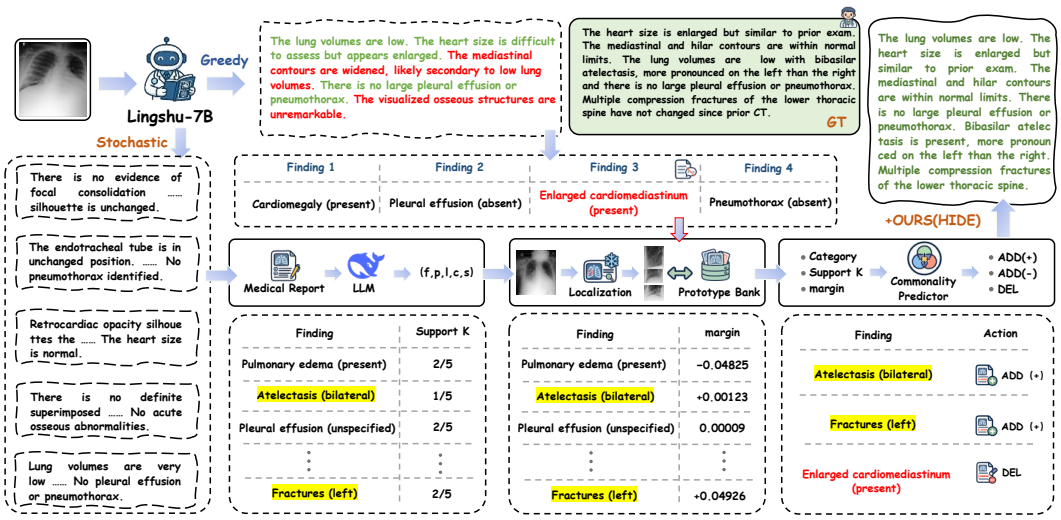

Figure 3: **Representative correction case.** HIDE adds recurrent findings from stochastic generations and removes an unstable claim from the initial report.

Table 2: **Comparison of correction methods on MIMIC-CXR with three report generators.** Ours (+HIDE) denotes the full model with Commonality Prediction (CP) and Prototype Bank (PB). Best result for each generator and metric is highlighted in bold, and the second-best is underlined.

Method	Radiology-specific Metrics					General NLG Metrics					
	CheXbert_{14} ↑	CheXbert_{5} ↑	RadGraph-F1 ↑	RaTEScore ↑	RadEval-BERT ↑	ROUGE-L ↑	BLEU-1 ↑	BLEU-4 ↑	METEOR ↑	CIDEr ↑	BERTScore ↑
Lingshu-7B	0.3439	0.3708	0.2021	0.5122	0.2075	0.2747	0.1737	0.0232	0.1683	0.9822	0.5037
+DeCo	0.3287	0.3410	0.2074	0.5110	0.1990	0.2802	0.1921	0.0255	0.1780	0.9793	0.5074
+ICD	0.3438	0.3704	0.2080	0.5154	0.2017	0.2759	0.1822	0.0251	0.1724	0.9870	0.5083
+VCD	0.3871	0.3978	0.2025	0.5000	0.1945	0.2694	0.1709	0.0209	0.1629	0.9419	0.4965
+CCD	0.3537	0.3785	0.1681	0.4992	0.1627	0.1806	0.1025	0.0139	0.1096	0.7624	0.4395
+CCS	0.3516	0.3651	0.1942	0.5016	0.2012	0.2622	0.1502	0.0160	0.1519	0.9181	0.4935
Ours (+HIDE)	**0.4591**	**0.4863**	**0.2196**	**0.5261**	**0.2131**	**0.2909**	**0.2240**	**0.0256**	**0.1947**	**1.0122**	**0.5137**
CheXAgent-2-3B	0.4104	0.4411	0.1289	0.4278	0.1639	0.1521	0.0507	0.0052	0.0784	0.4425	0.3180
+DeCo	0.4174	0.4725	0.1296	0.4530	0.1624	0.1579	0.0697	0.0043	0.0854	0.3813	0.2963
+ICD	0.2995	0.3562	0.0907	0.3653	0.1441	0.0972	0.0128	0.0010	0.0422	0.3372	0.2734
+VCD	0.4635	0.4893	0.1393	0.4288	0.1690	0.1573	0.0680	0.0070	0.0877	0.4564	0.3285
+CCD	0.4078	0.4616	0.1114	0.3980	0.1645	0.1378	0.0424	0.0038	0.0679	0.4147	0.3130
+CCS	0.4255	0.4484	0.1270	0.4236	0.1635	0.1461	0.0496	0.0045	0.0748	0.4345	0.3162
Ours (+HIDE)	**0.5220**	**0.5731**	**0.1722**	**0.4678**	**0.1709**	**0.1934**	**0.1240**	**0.0094**	**0.1208**	**0.4930**	**0.3457**
Med-R1-2B	0.2098	0.1357	0.0883	0.4383	0.1899	0.2057	0.2267	0.0037	0.1740	0.3808	0.3455
+DeCo	0.1980	0.1314	0.0896	0.4343	0.1863	0.1928	0.1919	0.0023	0.1560	0.3441	0.3296
+ICD	0.2391	0.1832	0.0997	0.4438	0.1875	0.2072	0.2242	0.0040	0.1803	0.4132	0.3501
+VCD	0.2204	0.1333	0.0883	0.4381	0.1869	0.2043	0.2229	0.0042	0.1773	0.3811	0.3523
+CCD	0.2474	0.2613	0.0998	0.4507	0.1795	**0.2228**	**0.2529**	**0.0077**	0.1839	0.4213	0.3552
+CCS	0.2147	0.1528	0.0791	0.4355	0.1774	0.1904	0.2116	0.0040	0.1625	0.3549	0.3603
Ours (+HIDE)	**0.3053**	**0.2965**	**0.1100**	**0.4541**	**0.1931**	0.2152	0.2258	0.0040	**0.1841**	**0.4255**	**0.3616**

Figure 3 presents a representative case in which additional stochastic samples uncover complementary imaging findings and identify an unreliable diagnostic statement in the initial report, illustrating the effectiveness of HIDE.

4.2 Comparison with Training-Free Correction Methods

Table 2 shows that HIDE ranks first on all three radiology-specific metrics across the three report generators. Compared with the decoding- or sequence-level baselines, its advantage comes from recovering omitted findings while filtering recurrent but unsupported candidates, thereby improving both report completeness and factuality. This is reflected in the CheXbert-F1-5 improvement for Med-R1-2B, from 0.1357 to 0.2965, and the consistent RadGraph-F1 gains across generators.

4.3 Performance With Limited Fitting Data

We further investigate the amount of fitting data required by HIDE. As shown in Figure 4, using only 0.1% of the fitting data yields positive mean relative gains on both MIMIC-CXR and CheXpert Plus, with gains generally increasing as the fitting-data fraction grows. When the fraction increases from 0.1% to 0.5%, CheXbert-F1-14 improves from 9.9% to 27.3% on MIMIC-CXR and shows a similar increase from 13.7% to 41.2% on CheXpert Plus. These results show that HIDE achieves substantial improvements with only a small fitting set, supporting correction based on image-grounded findings already available in frozen generators rather than retraining the generators.

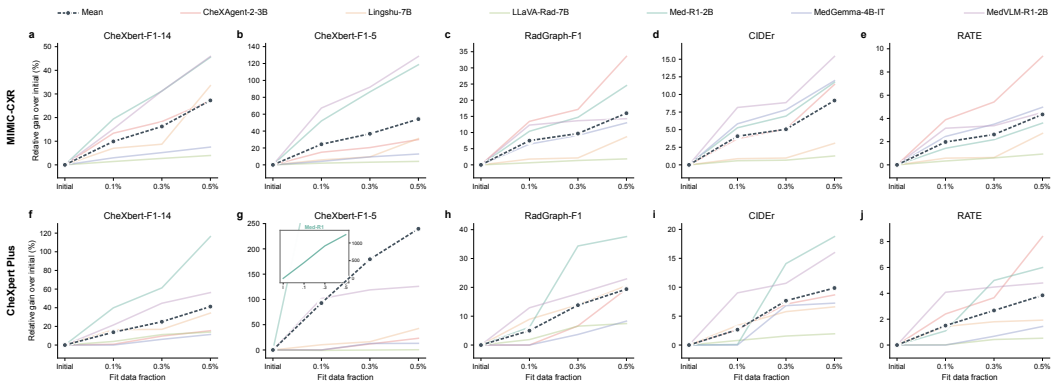

Figure 4: **Performance gains with limited fitting data.** Relative test-set gains across five metrics using 0.1%, 0.3%, and 0.5% of the fitting data on MIMIC-CXR (top) and CheXpert Plus (bottom).

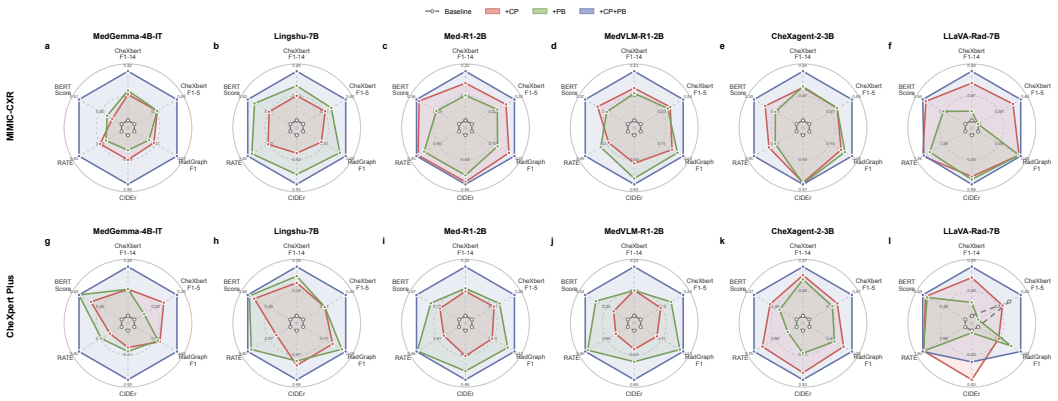

Figure 5: **Ablation results on MIMIC-CXR and CheXpert Plus.** The four configurations, Baseline, +CP, +PB and +CP+PB, are compared across six models and six evaluation metrics.

4.4 Ablation Study.

Figure 5 shows that both Commonality Prediction (CP) and the Prototype Bank (PB) improve report quality when used independently. CP provides proposal-side evidence from cross-sample recurrence, whereas PB supplies image-grounded verification. On MIMIC-CXR, CP improves CheXAgent-2-3B's CheXbert-F1-14 from 0.4104 to 0.4851 and RadGraph-F1 from 0.1289 to 0.1612, suggesting improved finding coverage and semantic alignment. PB improves Lingshu-7B's CheXbert-F1-14 from 0.3439 to 0.4251 and ROUGE-L from 0.2747 to 0.2882, likely through stronger visual verification. On CheXpert Plus, the combined configuration raises LLaVA-RAD-7B's CheXbert-F1-14 from 0.5113 to 0.5822 and RadGraph-F1 from 0.2230 to 0.2396. This combined setup yields the best performance across metrics, demonstrating the complementarity of CP and PB. Detailed results are provided in Appendix E.

4.5 Cross-Dataset Transferability

To assess cross-dataset transfer, we fit the corrector on either MIMIC-CXR or CheXpert Plus and apply it directly to IU-Xray while keeping all report generators frozen. As shown in Table 3, both source-domain adaptations improve the initial reports of all six models. Representative gains include $+40.0\%$ in CheXbert-F1-5 after MIMIC-CXR adaptation and $+11.3\%$ in RadGraph-F1 after CheXpert Plus adaptation. The improvements are particularly large for weaker initial reports, such as the $+109.8\%$ CheXbert-F1-5 gain for MedVLM-R1-2B and the $+47.5\%$ RadGraph-F1 gain for CheXagent-2-3B. These results indicate that HIDE transfers across datasets without target-domain fitting or generator updates, because it recovers and verifies candidate findings from the generator's stochastic outputs rather than relearning the target domain.

Table 3: **Direct transfer evaluation on IU-Xray after source-domain HIDE adaptation.** Here, M and C denote MIMIC-CXR and CheXpert Plus, respectively.

Adaptation	Radiology-specific Metrics					General NLG Metrics					
	CheXbert_{14} ↑	CheXbert_{5} ↑	RadGraph-F1 ↑	RaTEScore ↑	RadEval-BERT ↑	ROUGE-L ↑	B-1 ↑	B-4 ↑	METEOR ↑	CIDEr ↑	BERTScore ↑
Lingshu-7B	0.5417	0.3148	0.3423	0.6312	0.2631	0.3657	0.3445	0.0725	0.3081	1.5621	0.5714
+HIDE(M)	0.5594 (+3.3%)	0.3320 (+5.5%)	0.3538 (+3.4%)	0.6466 (+2.4%)	0.2688 (+2.2%)	0.3737 (+2.2%)	0.3483 (+1.1%)	0.0758 (+4.6%)	0.3192 (+3.6%)	1.6104 (+3.1%)	0.5762 (+0.8%)
+HIDE(C)	0.5544 (+2.3%)	0.3192 (+1.4%)	0.3494 (+2.1%)	0.6403 (+1.4%)	0.2687 (+2.1%)	0.3700 (+1.2%)	0.3468 (+0.7%)	0.0747 (+3.0%)	0.3181 (+3.2%)	1.5902 (+1.8%)	0.5736 (+0.4%)
MedGemma-4B-IT	0.6214	0.3908	0.3543	0.6531	0.2556	0.3708	0.3501	0.0674	0.3439	1.4383	0.5600
+HIDE(M)	0.6267 (+0.9%)	0.4444 (+13.7%)	0.3803 (+7.3%)	0.6975 (+6.8%)	0.2595 (+1.5%)	0.3890 (+4.9%)	0.3802 (+8.6%)	0.0732 (+8.6%)	0.3612 (+5.0%)	1.6365 (+13.8%)	0.5785 (+3.3%)
+HIDE(C)	0.6269 (+0.9%)	0.4563 (+16.8%)	0.3745 (+5.7%)	0.6775 (+3.7%)	0.2574 (+0.7%)	0.3839 (+3.5%)	0.3795 (+8.4%)	0.0726 (+7.7%)	0.3539 (+2.9%)	1.5442 (+7.4%)	0.5832 (+4.1%)
MedVLM-R1-2B	0.2737	0.0600	0.2026	0.5580	0.2208	0.2263	0.1171	0.0050	0.2113	0.7143	0.3776
+HIDE(M)	0.3405 (+24.4%)	0.1259 (+109.8%)	0.2107 (+4.0%)	0.5724 (+2.6%)	0.2260 (+2.4%)	0.2398 (+6.0%)	0.1524 (+30.1%)	0.0069 (+38.0%)	0.2151 (+1.8%)	0.7554 (+5.8%)	0.3867 (+2.4%)
+HIDE(C)	0.3288 (+20.1%)	0.0946 (+57.7%)	0.2083 (+2.8%)	0.5670 (+1.6%)	0.2298 (+4.1%)	0.2319 (+2.5%)	0.1398 (+19.4%)	0.0063 (+26.0%)	0.2132 (+0.9%)	0.7297 (+2.2%)	0.3842 (+1.7%)
LLaVA-RAD-7B	0.6166	0.4251	0.3119	0.6540	0.2487	0.3442	0.3481	0.0637	0.2948	1.4129	0.5831
+HIDE(M)	0.6384 (+3.5%)	0.4372 (+2.8%)	0.3219 (+3.2%)	0.6565 (+0.4%)	0.2496 (+0.4%)	0.3556 (+3.3%)	0.3599 (+3.4%)	0.0644 (+1.1%)	0.3053 (+3.6%)	1.4265 (+1.0%)	0.5842 (+0.2%)
+HIDE(C)	0.6382 (+3.5%)	0.4242 (-0.2%)	0.3236 (+3.8%)	0.6560 (+0.3%)	0.2488 (+0.0%)	0.3550 (+3.1%)	0.3588 (+3.1%)	0.0645 (+1.3%)	0.3044 (+3.3%)	1.4264 (+1.0%)	0.5837 (+0.1%)
Med-R1-2B	0.4305	0.0447	0.1844	0.5479	0.2138	0.2379	0.2205	0.0043	0.2223	0.6110	0.3859
+HIDE(M)	0.4656 (+8.2%)	0.0904 (+102.2%)	0.1976 (+7.2%)	0.5514 (+0.6%)	0.2143 (+0.2%)	0.2414 (+1.5%)	0.2245 (+1.8%)	0.0043 (+0.0%)	0.2303 (+3.6%)	0.6265 (+2.5%)	0.3963 (+2.7%)
+HIDE(C)	0.4801 (+11.5%)	0.0634 (+41.8%)	0.1957 (+6.1%)	0.5497 (+0.3%)	0.2148 (+0.5%)	0.2407 (+1.2%)	0.2286 (+3.7%)	0.0044 (+2.3%)	0.2349 (+5.7%)	0.6285 (+2.9%)	0.4009 (+3.9%)
CheXAgent-2-3B	0.4938	0.2923	0.1192	0.5135	0.2004	0.1615	0.1135	0.0156	0.1243	0.4227	0.2545
+HIDE(M)	0.5140 (+4.1%)	0.3100 (+6.1%)	0.1261 (+5.8%)	0.5174 (+0.8%)	0.2090 (+4.3%)	0.1663 (+3.0%)	0.1251 (+10.2%)	0.0170 (+9.0%)	0.1344 (+8.1%)	0.4419 (+4.5%)	0.2678 (+5.2%)
+HIDE(C)	0.5031 (+1.9%)	0.3071 (+5.1%)	0.1758 (+47.5%)	0.5763 (+12.2%)	0.2127 (+6.2%)	0.2069 (+28.1%)	0.1376 (+21.2%)	0.0165 (+5.8%)	0.1725 (+38.8%)	0.5878 (+39.1%)	0.3194 (+25.5%)
Mean Rel. Gain (M)	**+7.4%**	**+40.0%**	**+5.1%**	**+2.3%**	**+1.8%**	**+3.5%**	**+9.2%**	**+10.2%**	**+4.3%**	**+5.1%**	**+2.4%**
Mean Rel. Gain (C)	**+6.7%**	**+20.4%**	**+11.3%**	**+3.2%**	**+2.3%**	**+6.6%**	**+9.4%**	**+7.7%**	**+9.1%**	**+9.1%**	**+6.0%**

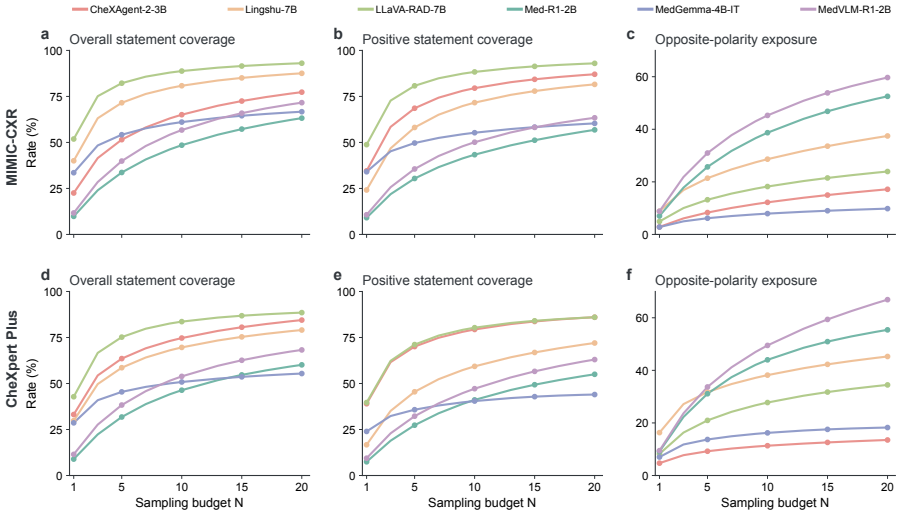

Figure 6: **Sampling coverage and polarity conflicts.** Overall and positive coverage, and opposite-polarity exposure, across six generators with sampling budgets $K = 1$ to 20.

4.6 Coverage Gains and Polarity Conflicts

We conducted an independent sampling-coverage analysis to characterize the candidate statements available across repeated generations from the same model and their potential conflicts (Figure 6). Increasing the sampling budget from 1 to 20 raises overall and positive-statement coverage for all six models, with overall gains of 33.1%–59.9% on MIMIC-CXR and 26.8%–56.8% on CheXpert Plus. The gains diminish at larger budgets. Opposite-polarity exposure increases concurrently, showing that repeated sampling broadens the candidate pool while introducing conflicting alternatives, thereby motivating cross-sample support and localized visual verification before report editing.

5 Conclusion

We challenge the need to retrain the report generator to recover omitted findings. A frozen medical MLLM can produce complementary, image-supported findings across stochastic decoding runs. Building on this observation, we introduce HIDE, a training-free framework that extracts finding-level candidates, combines cross-sample recurrence with localized visual evidence, and applies source-constrained minimal edits. Experiments across six medical MLLMs show that HIDE improves report quality and achieves state-of-the-art performance against existing methods for improving MLLM-based radiology report generation. These demonstrate that stochastic decoding provides a practical source of recoverable clinical information for reliable report correction without updating the generator. *Limitations.* HIDE relies on a prototype bank constructed from Chest ImaGenome, so its visual verification is constrained by the coverage of findings and anatomical regions represented in the dataset may benefit from broader coverage of findings and anatomical regions.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, including sentence rephrasing, grammar correction, and improving readability of manuscript text. Generative AI was also used for retrieval and discovery (e.g., finding related work). Additionally, no other generative AI-assisted tasks are involved. We have reviewed all AI-assisted content thoroughly. All outputs produced by generative AI were manually checked and revised by the authors to ensure accuracy, consistency and scientific rigor. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All datasets used in this work, including MIMIC-CXR, CheXpert Plus, IU-Xray and Chest ImaGenome, are publicly available de-identified medical imaging datasets. All experimental procedures strictly follow the corresponding data-usage agreements for each dataset. No raw information that could identify patient identities is processed in our research pipeline. In addition, there are no conflicts of interest in this study.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide detailed descriptions of dataset selection, preprocessing pipelines, evaluation metrics and implementation configurations in the main text and appendices. Appendix A documents model-specific hyper-parameters and sampling budget K; Appendix B describes the filtering procedure for the prototype bank; Appendix C presents prompt configurations for clinical-finding extraction; Appendix D records branch-wise thresholds of the commonality predictor; Appendix E contains full ablation results for each model. We adopt six publicly-available medical multimodal large-language-model checkpoints, and dataset preprocessing strictly follows standard protocols established in prior work. Upon acceptance of this paper, we will release the complete source code to enable researchers to fully reproduce all quantitative experiments and case analyses presented in this paper.

REFERENCES

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

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. *arXiv preprint arXiv:2405.19538*, 2024.

Zailong Chen, Peng Gao, Johan Barthelemy, Luping Zhou, and Lei Wang. Analyzing and enhancing visual learning in llm-based radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9327–9336, June 2026a.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, et al. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

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

Sheng Cheng and Devika Subramanian. Rethinking radiology report generation: From narrative flow to topic-guided findings. In *International Conference on Learning Representations*, 2026.

Yinlam Chow, Guy Tennenholtz, Izzeddin Gur, Vincent Zhuang, Bo Dai, Aviral Kumar, Rishabh Agarwal, Sridhar Thiagarajan, Craig Boutilier, and Aleksandra Faust. Inference-aware fine-tuning for best-of-n sampling in large language models. In *International Conference on Learning Representations*, 2025.

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

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2015.

Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014.

Debanjan Goswami, Ronast Subedi, and Shayok Chakraborty. Medivlm: A vision language model for radiology report generation from medical images. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10287–10304. Association for Computational Linguistics, 2025.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Lili Huang, Yiming Cao, Pengcheng Jia, Chenglong Li, Jin Tang, and Chuanfu Li. Knowledge-guided cross-modal alignment and progressive fusion for chest x-ray report generation. *IEEE Transactions on Multimedia*, 27:557–567, 2025.

Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.

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

Manav Nitin Kapadnis, Sohan Patnaik, Abhilash Nandy, Sourjyadip Ray, Pawan Goyal, and Debdoot Sheet. SERPENT-vlm: Self-refining radiology report generation using vision language models. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pp. 283–291, Mexico City, Mexico, 2024. Association for Computational Linguistics.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE transactions on medical imaging*, 2026.

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.

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

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10348–10359, 2025.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis 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, Online, 2021. Association for Computational Linguistics.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 337–347. Springer, 2025.

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

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

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Harshita Sharma, Valentina Salvatelli, Shaury Srivastav, Kenza Bouzid, Shruthi Bannur, Daniel C. Castro, Maximilian Ilse, Sam Bond-Taylor, Mercy Prasanna Ranjit, Fabian Falck, Fernando Pérez-García, Anton Schwaighofer, Hannah Richardson, Maria Teodora Wetscherek, Stephanie Hyland, and Javier Alvarez-Valle. Maira-seg: Enhancing radiology report generation with segmentation-aware multimodal large language models. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 941–960. PMLR, 2025.

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.

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

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

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, 2025a.

Chenyu Wang, Weichao Zhou, Shantanu Ghosh, Kayhan Batmanghelich, and Wenchao Li. Semantic consistency-based uncertainty quantification for factuality in radiology report generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1739–1754. Association for Computational Linguistics, 2025b.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853, 2024.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset. *PhysioNet*, 2021.

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

Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. Radiology report generation via multi-objective preference optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8664–8672, 2025.

Justin Xu, Xi Zhang, Javid Abderezaei, Julie Bauml, Roger Boodoo, Fatemeh Haghighi, Ali Ganjizadeh, Eric Brattain, Dave Van Veen, Zaiqiao Meng, David W. Eyre, and Jean-Benoit Delbrouck. Radeval: A framework for radiology text evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 546–557, 2025.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Can Hu, Siwei Pan, Jie Tan, Tingyang Xu, Deli Zhao, Hao Zhang, and Yu Rong. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Sixing Yan, Ziao Wang, Kejing Yin, William Kwok-wai Cheung, Ka Chun Cheung, and Simon See. Learning self-critiquing mechanisms for region-guided chest x-ray report generation. In *International Conference on Learning Representations*, 2026.

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

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(3108), 2025.

Serena Zhang, Sraavya Sambara, Oishi Banerjee, Julian N. Acosta, L. John Fahrner, and Pranav Rajpurkar. Radflag: A black-box hallucination detection method for medical vision language models. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 1087–1103. PMLR, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Xi Zhang, Yingshu Li, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. Ccs: Clinical consensus selection for radiology report generation. *arXiv preprint arXiv:2605.30131*, 2026a.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. CCD: Mitigating hallucinations in radiology MLLMs via clinical contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35174–35200, San Diego, California, United States, 2026b. Association for Computational Linguistics.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15004–15019, 2024.

A Model-Specific Selection of the Number of Additional Samples

The number of additional stochastic samples K controls how many alternative generations are available for recovering findings omitted from the initial draft. Because different report generators exhibit different levels of sampling diversity, we determine K separately for each generator prior to evaluation.

A.1 Pilot Cohort and Stochastic Sampling

We construct an independent pilot cohort of 500 patient-disjoint MIMIC-CXR studies that does not overlap with any fitting set or final evaluation set. We evaluate six frozen report generators: MedGemma-4B-IT, Lingshu-7B, Med-R1-2B, MedVLM-R1-2B, CheXagent-2-3B, and LLaVA-Rad-7B.

For each generator and study, we first obtain an initial draft $r_i^{(0)}$ using deterministic decoding. We then generate 12 additional stochastic samples,

$$r_i^{(1)}, \ldots, r_i^{(12)},$$

using temperature 0.8, top-$p = 0.95$, and independent random seeds. The reference report is never provided to the report generator during this process. The experiment therefore measures only what additional findings can be exposed by repeated decoding from the same frozen generator.

A.2 Metric Definitions

Standard report-level metrics do not directly measure the specific capability of interest here: whether additional stochastic samples recover findings that were missed by the initial draft. We therefore define a set of **Addition metrics** for this purpose.

For study i and generator m, let T_i denote the set of positive findings extracted from the reference report, and let $B_i^{(m)}$ denote the positive findings contained in the deterministic initial draft. The findings missed by the initial draft are

$$M_i^{(m)} = T_i \setminus B_i^{(m)}.$$

For a given sampling budget K, let $S_{i,j}^{(m)}$ denote the positive findings extracted from the j-th stochastic sample. We first take the union of findings observed in the first K samples and then remove findings already present in the initial draft:

$$A_{i,K}^{(m)} = \left(\bigcup_{j=1}^{K} S_{i,j}^{(m)} \right) \setminus B_i^{(m)}.$$

Thus, $A_{i,K}^{(m)}$ contains only **new candidate findings exposed by additional stochastic sampling**.

We then define

$$\mathrm{TP}_K^{(m)} = \sum_i \left| A_{i,K}^{(m)} \cap M_i^{(m)} \right|,$$

$$\mathrm{FP}_K^{(m)} = \sum_i \left| A_{i,K}^{(m)} \setminus T_i \right|,$$

and

$$\mathrm{FN}_K^{(m)} = \sum_i \left| M_i^{(m)} \setminus A_{i,K}^{(m)} \right|.$$

Accordingly,

$$P_K^{\mathrm{add},(m)} = \frac{\mathrm{TP}_K^{(m)}}{\mathrm{TP}_K^{(m)} + \mathrm{FP}_K^{(m)}},$$

$$R_K^{\mathrm{add},(m)} = \frac{\mathrm{TP}_K^{(m)}}{\mathrm{TP}_K^{(m)} + \mathrm{FN}_K^{(m)}},$$

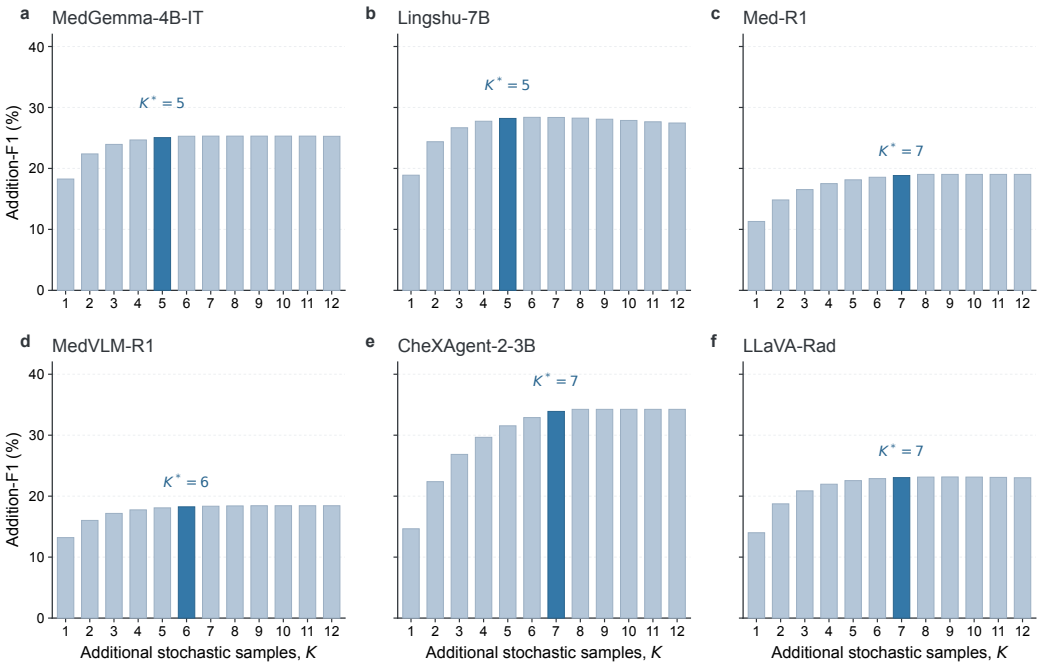

Figure A1: **Model-specific selection of stochastic sampling operating points.** Addition-F1 as a function of the number of additional stochastic samples K. The highlighted point indicates the selected operating point K^*, defined as the smallest sampling budget achieving at least 99% of the maximum Addition-F1.

and

$$F1_K^{\text{add},(m)} = \frac{2P_K^{\text{add},(m)} R_K^{\text{add},(m)}}{P_K^{\text{add},(m)} + R_K^{\text{add},(m)}}.$$

Addition-Precision measures the fraction of newly exposed candidate findings that are supported by the reference report.

Addition-Recall measures the fraction of findings missed by the initial draft that are recovered from the additional stochastic samples.

Addition-F1 summarizes this recovery–noise trade-off and is used only to select K.

All generated and reference reports are mapped to the same 14-category CheXbert finding space before computing these metrics. Repeated mentions of the same finding within a report are counted only once.

A.3 Model-Specific K Selection

For each generator, we evaluate candidate budgets:

$$K \in \{1, 2, \ldots, 12\}.$$

The optimal recovery performance is:

$$F_{\max} = \max_K F1_{\text{add}}(K).$$

We select the smallest sampling budget that reaches 99% of the maximum observed Addition-F1:

$$K^* = \min \{K : F1_{\text{add}}(K) \geq 0.99 F_{\max}\}.$$

This criterion selects a compact sampling budget while preserving near-maximal recovery of omitted findings. The selected K^* values are fixed before final evaluation and remain unchanged throughout all subsequent experiments.

As shown in Fig. A1, Addition-F1 exhibits clear model-specific saturation. The resulting operating points are $K = 5$ for MedGemma-4B-IT, $K = 5$ for Lingshu-7B, $K = 7$ for Med-R1-2B, $K = 6$ for MedVLM-R1-2B, $K = 7$ for CheXagent-2-3B, and $K = 7$ for LLaVA-Rad-7B. These values are fixed before the final evaluation and are used throughout all subsequent experiments.

B Prototype Bank Construction

B.1 Data Source and Patient-Level Split

The prototype bank is constructed using Chest ImaGenome with linked MIMIC-CXR-JPG images. Chest ImaGenome provides scene graph annotations describing radiological findings, attributes, and anatomical relations, together with anatomical bounding boxes that associate findings with localized image regions. To avoid information leakage, patients appearing in validation, threshold selection, calibration, or test sets are excluded before prototype construction.

The prototype bank is organized by finding category, anatomical region, imaging projection, and prototype role. We consider twelve target findings: atelectasis, cardiomegaly, consolidation, pulmonary edema, enlarged cardiomediastinum, fracture, lung lesion, lung opacity, pleural effusion, pleural other, pneumonia, and pneumothorax. Each prototype entry corresponds to a specific finding, anatomical region, projection view, and role, where the role indicates whether the reference image provides positive evidence, normal evidence, or confounding visual evidence for the target finding.

B.2 Prototype Roles

For each finding-specific entry, we construct three types of visual references: positive, normal, and confounder prototypes (see Table B.1).

Table B.1: Definitions of prototype roles.

Role	**Definition**
Positive	The target finding is explicitly present in the corresponding anatomical region.
Normal	The target finding is explicitly absent without an admitted visually confounding abnormality.
Confounder	The target finding is explicitly absent while a visually similar alternative abnormality is explicitly present.

Positive prototypes provide visual references of the target finding. Normal and confounder prototypes provide counterevidence by distinguishing the target finding from both normal appearances and clinically similar abnormalities.

Prototype role assignment follows a frozen finding-specific mapping from Chest ImaGenome annotations. Non-mention of a finding is not treated as evidence of absence and cannot be used to construct normal or confounder prototypes. Uncertain or conflicting annotations are excluded rather than assigned to any role.

B.3 Finding-Specific Prototype Assignment Rules

To construct finding-specific prototype references, we define a frozen mapping between Chest ImaGenome attributes and the target findings (see Table B.2).

C Prompt Details

We provide the prompts used for report generation and structured finding extraction. During inference, the report generator receives only the chest X-ray image and the fixed generation prompt. The extraction model receives only the generated report text and the extraction prompt, without access to images, reference reports, prototype information, or correction decisions.

Table B.2: Finding-specific prototype assignment rules.

Finding	Positive attributes	Sufficient negative evidence	Allowed confounder alternatives
Atelectasis	atelectasis; linear/patchy atelectasis; lobar/segmental collapse	atelectasis=no	consolidation; pneumonia; pulmonary edema; pleural/parenchymal scarring; low lung volumes
Cardiomegaly	enlarged cardiac silhouette	enlarged cardiac silhouette=no on cardiac silhouette	pericardial effusion; enlarged cardiomediastinum; tortuous aorta
Consolidation	consolidation; airspace opacity	consolidation=no or airspace opacity=no	atelectasis; pneumonia; pulmonary edema; generic lung opacity; scarring
Pulmonary edema	pulmonary edema/hazy opacity; vascular congestion; vascular redistribution	complete bilateral lung or bilateral hilar negative coverage	bilateral/multifocal pneumonia; consolidative opacity; atelectatic opacity
Enlarged cardiomediastinum	mediastinal widening; superior mediastinal mass/enlargement	complete mediastinal negative coverage	cardiomegaly; tortuous aorta; pericardial effusion; mediastinal displacement/hernia
Fracture	rib fracture; clavicle fracture; spinal fracture	complete spine or clavicle negative coverage	bone lesion; spinal degenerative change (spine only)
Lung lesion	lung lesion; mass/nodule; multiple masses/nodules; calcified nodule	lung lesion=no	nipple/breast shadow; focal pleural/parenchymal scarring; focal nonspecific opacity/consolidation
Lung opacity	lung opacity; infiltration	lung opacity=no	pleural effusion; pleural thickening/scarring; nipple shadow; explicit low lung volumes
Pleural effusion	pleural effusion; costophrenic angle blunting	pleural effusion=no	pleural thickening/scarring; basal atelectasis; basal consolidation/opacity
Pleural other	pleural/parenchymal scarring	no explicit sufficient negative rule is defined for this finding	pleural effusion; costophrenic angle blunting; atelectasis
Pneumonia	pneumonia	pneumonia=no	atelectasis; non-pneumonic consolidation; pulmonary edema; scarring; generic lung opacity
Pneumothorax	pneumothorax; hydropneumothorax	pneumothorax=no	cyst/bullae; skin fold; apical pleural scar

C.1 Report Generation Prompt

The following prompt is used to generate the initial draft and the additional stochastic samples:

> You are an expert radiologist tasked with interpreting the provided chest X-ray image. Generate a concise diagnostic radiology report based solely on observations supported by the image. Return the report as plain text using exactly the following format: `Findings: ...`. Do not include an Impression, Conclusion, Recommendations, any section heading other than Findings, or any additional explanation.

C.2 Structured Finding Extraction Prompt

The following prompt is used with DeepSeek-V4-Pro to extract structured findings from the initial draft and additional stochastic samples:

```
You are a radiology report entity extractor. Return one JSON object only with exactly this
     outer structure: {"batch_id":"copy exactly","results":[{"report_key":"copy exactly
    ","entities":[...]}]}. Process every report independently and return one result for
    every supplied report_key. Each entity must contain exactly these five keys:
    canonical_finding, polarity, laterality, target_clause, source_sentence.

The only supported canonical_finding values are: atelectasis, cardiomegaly, consolidation,
     enlarged cardiomediastinum, fracture, lung lesion, lung opacity, pleural effusion,
    pleural other, pneumonia, pneumothorax, pulmonary edema.

Do not force unsupported observations into these categories. Explicit interstitial lung
    disease, pulmonary fibrosis, or definite chronic interstitial disease must not be
    extracted and must not be reassigned to lung opacity. Interstitial markings or
    opacities may map to lung opacity only when an opacity or radiographic marking is
    itself explicitly described as a current finding; findings explicitly attributed to
    edema or congestion map to pulmonary edema. Fracture means bone fracture only, never
    a broken wire, lead, catheter, or other device. Pleural other is limited to explicit
    pleural plaques, pleural calcification, or pleural thickening. Enlarged heart or
    cardiac silhouette maps to cardiomegaly; widened or enlarged mediastinum maps to
    enlarged cardiomediastinum.
```

```
Extract both coarse and fine findings when both are explicit. In "opacity may represent
     pneumonia," lung opacity is present while pneumonia is uncertain; do not infer a
     coarse finding when no coarse term is written. Use JSON null, never the string "null
     ". Keep distinct mentions from different sentences or report sections, but do not
     duplicate the same mention or expand generic normal statements into unstated negative
      findings.

Extract every explicit current-examination mention of one of the supported findings,
     including present, absent, and uncertain mentions. Do not extract history, indication
     , devices, procedures, generic normal statements, or unsupported diagnoses. A clear
     negative statement such as "no pleural effusion" or "without pneumothorax" MUST be
     extracted as polarity absent. A phrase such as "possible pneumonia" or "may represent
      atelectasis" MUST be extracted as polarity uncertain. Retain separate mentions in
     separate sentences or report sections; do not merge distinct occurrences. Return
     entities in report order, and preserve left-to-right order within a sentence.

CRITICAL VERBATIM-COPY CONTRACT:
1. source_sentence MUST be copied character-for-character from one contiguous span of
     report_text. Preserve its capitalization, punctuation, spacing, spelling, and wording
     . Never remove a leading "Findings:" when it belongs to the copied span. If sentence
     boundaries are uncertain, copy a larger contiguous span or the complete report_text.
2. target_clause MUST also be copied character-for-character from one contiguous span of
     report_text. It may equal source_sentence. Never paraphrase, normalize capitalization
     , reconstruct an implied phrase, insert omitted shared words, reorder words, or use
     ellipses.
3. For coordinated findings with a shared modifier, copy the entire literal coordinated
     phrase whenever the smaller semantic phrase is not contiguous. Example: from "There
     is no pleural effusion or pneumothorax.", the pneumothorax entity MUST NOT use "no
     pneumothorax" because that string is absent. Use the exact contiguous target "no
     pleural effusion or pneumothorax" (or the full exact sentence) for both supported
     mentions. Likewise, never turn "no focal consolidation, pleural effusion, or
     pneumothorax" into a reconstructed "no pleural effusion".
4. Before returning JSON, verify for every entity that target_clause occurs exactly in its
      own report_text and source_sentence occurs exactly in its own report_text. If an
     exact span cannot be copied, omit that entity rather than inventing a span.

polarity must be exactly present, absent, or uncertain. laterality must be exactly left,
     right, bilateral, or JSON null. Words such as diffuse, retrocardiac, central,
     bibasilar, and the string "null" are never valid laterality values. Use laterality
     only when explicitly stated or grammatically shared. Do not add severity, anatomy,
     explanations, confidence scores, action labels, prototype roles, or markdown. Output
     JSON only.
```

D BRANCH-SPECIFIC COMMONALITY THRESHOLDS

In the HIDE framework, candidate findings extracted across stochastic decodes serve distinct diagnostic objectives across different editing branches (ADD^+, ADD^-, and DEL). Specifically, ADD^+ recovers image-supported positive or uncertain findings omitted from the initial greedy draft, ADD^- identifies explicitly absent (normal) findings to enhance clinical completeness, and DEL prunes unsupported or hallucinated claims from the initial draft. Because the distributions of proposal recurrence and localized visual support differ substantially across these branches and base report generators, branch-specific decision thresholds are calibrated to govern candidate authorization.

Formally, for each branch $b \in \{\mathrm{ADD}^+, \mathrm{ADD}^-, \mathrm{DEL}\}$, the commonality predictor $T_b(\phi_b(g))$ produces an authorization score $q_b(g) \in [0, 1]$ conditioned on the finding category, the cross-sample support count $k_b(g)$, and the prototype margin $m(g)$. A candidate finding is authorized for source-constrained minimal editing if $q_b(g) \geq \tau_b$. To preserve high clinical precision and prevent aggressive alterations, the branch-specific thresholds τ_b are calibrated via a grid search on the held-out 0.5% fitting split of MIMIC-CXR and CheXpert Plus, targeting the optimal balance between clinical observation recall (CheXbert-F_1) and factual consistency.

Table D.1 reports the calibrated thresholds τ_{ADD^+}, τ_{ADD^-}, and τ_{DEL} across all six frozen medical MLLMs on both datasets. We observe two consistent patterns across generators:

- **Conservative deletion and absence thresholds (τ_{ADD^-} and τ_{DEL}):** Both thresholds are consistently set high (typically exceeding 0.75), reflecting a conservative criterion to avoid mistakenly removing true clinical observations or asserting unsubstantiated negations.

- **Flexible positive recall (τ_{ADD^+}):** The threshold for positive additions remains substantially lower (typically between 0.04 and 0.50), effectively empowering stochastic decoding to capture subtle or low-probability abnormalities that were omitted by the greedy baseline.

Table D.1: Branch-specific thresholds (τ_{ADD^+}, τ_{ADD^-}, and τ_{DEL}) for commonality prediction across six frozen medical MLLMs on MIMIC-CXR and CheXpert Plus.

Model	MIMIC-CXR			CheXpert Plus		
	τ_{ADD^+}	τ_{ADD^-}	τ_{DEL}	τ_{ADD^+}	τ_{ADD^-}	τ_{DEL}
Lingshu-7B	0.195	0.748	0.923	0.200	0.836	0.800
MedGemma-4B-IT	0.099	0.891	0.972	0.042	0.894	0.925
MedVLM-R1-2B	0.182	0.918	0.914	0.091	0.980	0.892
LLaVA-RAD-7B	0.440	0.837	0.909	0.500	0.842	0.833
Med-R1	0.235	0.905	0.545	0.200	0.911	0.286
CheXAgent-2-3B	0.188	0.789	0.733	0.231	0.813	0.786

E Ablation Analysis

This appendix provides a detailed analysis of the ablation results shown in Figure 5. We compare four configurations, namely Baseline, +CP, +PB, and +CP+PB, across six report generators on MIMIC-CXR and CheXpert Plus. Multi-sampling is kept fixed across all configurations.

Overall effect. Both CP and PB improve the radiology-specific metrics for most generators, while the combined configuration generally achieves the strongest performance. For example, on MIMIC-CXR, CheXAgent-2-3B improves from 0.4104 to 0.4851 in CheXbert-F1-14 with CP alone and to 0.4873 with PB alone. Combining both modules further increases the score to 0.5220. The same trend is observed on CheXpert Plus, where the combined configuration reaches 0.5597, compared with 0.4851 for the baseline. CheXbert-F1-5 shows a similar improvement, increasing from 0.4411 to 0.5731 on MIMIC-CXR and from 0.4531 to 0.5578 on CheXpert Plus.

Complementary roles of CP and PB. CP provides proposal-side information from cross-sample recurrence. For CheXAgent-2-3B, it improves RadGraph-F1 from 0.1289 to 0.1612 on MIMIC-CXR and from 0.2089 to 0.2415 on CheXpert Plus. PB provides image-grounded verification and is particularly effective for models such as Lingshu-7B. On MIMIC-CXR, PB increases its CheXbert-F1-14 from 0.3439 to 0.4251, while on CheXpert Plus it increases the score from 0.2807 to 0.3584. These results indicate that CP and PB improve report generation through different but complementary sources of evidence.

Effect on language quality. The improvements are not limited to clinical metrics. On MIMIC-CXR, the combined configuration increases Lingshu-7B's ROUGE-L from 0.2747 to 0.2909 and CIDEr from 0.9822 to 1.0122. On CheXpert Plus, it improves LLaVA-RAD-7B's ROUGE-L from 0.2662 to 0.2765 and BERTScore from 0.4806 to 0.4884. Although an individual component occasionally achieves the best value for a specific NLG metric, the combined configuration is generally best or tied best across the radiology-specific metrics and most NLG metrics.

Overall, the ablation results support the complementary design of CP and PB. CP supplies recurrence-based candidate evidence, whereas PB verifies candidates against localized visual information. Their combination therefore improves both clinical correctness and report-level language quality across different generators and datasets.

F Scope of the Training-Free Claim

The term *training-free* in HIDE is defined with respect to the report generator. Let G_θ denote one of the released medical MLLMs. In all experiments, θ is loaded from the released checkpoint and remains fixed. No gradients, teacher-forced report targets, preference updates, LoRA/adapters, or image–text fine-tuning are applied to G_θ. The tokenizer, generation prompt, decoding parameters, and RAD-DINO image encoder are also fixed.

Table E.1: Detailed ablation results of Commonality Prediction (CP) and the Prototype Bank (PB) on MIMIC-CXR and CheXpert Plus. Baseline denotes the original report generator, while +CP, +PB, and +CP+PB introduce the two components individually and jointly. Multi-sampling is fixed across all variants. Bold values indicate the best result for each generator and metric within each dataset.

Model	Radiology-specific Metrics					NLG Metrics					
	CheXbert_{14} ↑	CheXbert_{5} ↑	RadGraph-F1 ↑	RaTE ↑	RadEval-BERT ↑	R-L ↑	B-1 ↑	B-4 ↑	METEOR ↑	CIDEr ↑	BERTScore ↑
MIMIC-CXR											
Lingshu-7B	0.3439	0.3708	0.2021	0.5122	0.2075	0.2747	0.1737	0.0232	0.1683	0.9822	0.5037
+ CP	0.4018	0.4306	0.2096	0.5173	0.2110	0.2814	0.1902	0.0239	0.1764	0.9978	0.5087
+ PB	0.4251	0.4471	0.2172	0.5233	0.2121	0.2882	0.2142	0.0253	0.1894	1.0093	0.5121
+ CP+PB	**0.4591**	**0.4863**	**0.2196**	**0.5261**	**0.2131**	**0.2909**	**0.2240**	**0.0256**	**0.1947**	**1.0122**	**0.5137**
MedGemma-4B-IT	0.4783	0.4584	0.2008	0.5252	0.2075	0.2555	0.2288	0.0279	0.2030	0.8107	0.4936
+ CP	0.4973	0.4900	0.2128	0.5383	0.2087	0.2681	0.2620	0.0306	0.2174	0.8597	0.4970
+ PB	0.5002	0.4900	0.2098	0.5332	0.2087	0.2644	0.2467	0.0297	0.2114	0.8462	0.4989
+ CP+PB	**0.5143**	**0.5173**	**0.2269**	**0.5513**	**0.2097**	**0.2828**	**0.2813**	**0.0320**	**0.2299**	**0.9076**	**0.5092**
MedVLM-R1-2B	0.2170	0.1386	0.1044	0.4419	0.1930	0.2148	0.2214	**0.0057**	0.1835	0.4164	0.3600
+ CP	0.2824	0.2623	0.1154	0.4533	0.1955	0.2191	0.2220	**0.0057**	0.1911	0.4453	0.3683
+ PB	0.2719	0.2521	0.1173	**0.4594**	0.1957	0.2197	0.2169	0.0056	**0.1933**	0.4573	0.3658
+ CP+PB	**0.3165**	**0.3165**	**0.1193**	**0.4618**	**0.1974**	**0.2206**	**0.2220**	**0.0057**	0.1918	**0.4805**	**0.3717**
LLaVA-RAD-7B	0.5634	0.5741	0.2747	0.5494	0.2543	0.3248	0.2913	0.0643	0.2598	1.2695	0.5412
+ CP	0.5804	0.5934	0.2796	0.5537	0.2574	0.3302	0.3031	0.0653	0.2645	1.2859	0.5437
+ PB	0.5676	0.5739	0.2796	0.5540	0.2552	0.3302	0.3006	0.0653	0.2648	1.2832	0.5426
+ CP+PB	**0.5858**	**0.5977**	**0.2798**	**0.5545**	**0.2576**	**0.3303**	**0.3036**	**0.0655**	**0.2647**	**1.2855**	**0.5439**
Med-R1-2B	0.2098	0.1357	0.0883	0.4383	0.1899	0.2057	**0.2267**	0.0037	0.1740	0.3808	0.3455
+ CP	0.2822	0.2646	0.1072	0.4533	0.1910	0.2144	0.2256	0.0038	0.1836	0.4239	0.3606
+ PB	0.2583	0.2318	0.1014	0.4512	0.1910	0.2109	0.2242	0.0038	0.1823	0.4175	0.3541
+ CP+PB	**0.3053**	**0.2965**	**0.1100**	**0.4541**	**0.1931**	**0.2152**	0.2258	**0.0040**	**0.1841**	**0.4255**	**0.3616**
CheXAgent-2-3B	0.4104	0.4411	0.1289	0.4278	0.1639	0.1521	0.0507	0.0052	0.0784	0.4425	0.3180
+ CP	0.4851	0.5274	0.1612	0.4664	0.1646	0.1873	0.1071	0.0085	0.1122	0.4753	0.3384
+ PB	0.4873	0.5269	0.1651	0.4672	0.1642	0.1879	0.1183	0.0087	0.1160	0.4678	0.3318
+ CP+PB	**0.5220**	**0.5731**	**0.1722**	**0.4678**	**0.1709**	**0.1934**	**0.1240**	**0.0094**	**0.1208**	**0.4930**	**0.3457**
CheXpert Plus											
Lingshu-7B	0.2807	0.2970	0.1506	0.4653	0.1979	0.2210	0.1272	0.0103	0.1279	0.7440	0.4438
+ CP	0.3453	0.3611	0.1720	0.4716	0.1956	0.2304	0.1519	0.0104	0.1431	0.7601	0.4549
+ PB	0.3584	0.3611	0.1787	0.4708	0.1992	**0.2352**	0.1568	0.0107	0.1429	0.7890	0.4564
+ CP+PB	**0.3769**	**0.4228**	**0.1815**	**0.4742**	**0.2000**	**0.2352**	**0.1574**	**0.0109**	**0.1434**	**0.7931**	**0.4569**
MedGemma-4B-IT	0.4372	0.3368	0.1845	0.5070	0.2193	0.2391	0.2097	0.0180	0.1773	0.7665	0.4510
+ CP	0.4640	0.3673	0.1937	0.5095	0.2170	0.2438	0.2133	0.0179	0.1863	0.7817	0.4528
+ PB	0.4640	0.3469	0.1930	0.5101	0.2195	0.2431	0.2104	**0.0181**	0.1850	0.7929	**0.4535**
+ CP+PB	**0.4867**	**0.3810**	**0.1998**	**0.5143**	**0.2193**	**0.2478**	**0.2246**	**0.0181**	**0.1916**	**0.8221**	**0.4535**
MedVLM-R1-2B	0.1973	0.1379	0.0990	0.4371	0.2020	0.1979	**0.1774**	0.0023	0.1501	0.4038	0.3309
+ CP	0.2551	0.2200	0.1077	0.4451	0.2105	0.1998	0.1756	0.0024	0.1547	0.4220	0.3341
+ PB	0.2545	0.2623	0.1199	0.4504	0.2106	0.2019	0.1653	0.0027	0.1622	0.4647	0.3381
+ CP+PB	**0.3083**	**0.3115**	**0.1217**	**0.4581**	**0.2143**	**0.2023**	**0.1774**	**0.0030**	**0.1631**	**0.4683**	**0.3405**
LLaVA-RAD-7B	0.5113	0.5161	0.2230	0.5212	0.2251	0.2662	0.1976	0.0204	0.1740	0.9817	0.4806
+ CP	0.5664	0.5147	0.2310	0.5256	0.2316	0.2749	0.2053	0.0186	0.1779	1.0004	0.4879
+ PB	0.5307	0.5096	0.2359	0.5214	0.2304	0.2761	**0.2071**	**0.0204**	0.1779	1.0005	0.4876
+ CP+PB	**0.5822**	**0.5185**	**0.2396**	**0.5240**	**0.2331**	**0.2765**	**0.2071**	**0.0204**	**0.1780**	**1.0008**	**0.4884**
Med-R1-2B	0.1468	0.0250	0.0839	0.4269	0.2104	0.1911	0.1901	0.0015	0.1518	0.3791	0.3423
+ CP	0.2318	0.1856	0.0990	0.4401	0.2136	0.1969	0.1881	0.0018	0.1580	0.4045	0.3530
+ PB	0.2424	0.2254	0.1105	0.4480	0.2129	0.2013	0.1791	0.0015	0.1652	0.4488	0.3581
+ CP+PB	**0.3175**	**0.3333**	**0.1155**	**0.4525**	**0.2177**	**0.2056**	**0.1911**	**0.0024**	**0.1655**	**0.4502**	**0.3661**
CheXAgent-2-3B	0.4851	0.4531	0.2089	0.4855	0.2094	0.1931	0.1014	0.0078	0.1060	0.6353	0.3222
+ CP	0.5467	0.5217	0.2415	0.5204	0.2126	0.2295	0.1359	0.0085	0.1344	0.6794	0.3495
+ PB	0.5402	0.5096	0.2329	0.5043	0.2149	0.2123	0.1322	0.0092	0.1277	0.6467	0.3397
+ CP+PB	**0.5597**	**0.5578**	**0.2495**	**0.5262**	**0.2156**	**0.2380**	**0.1532**	**0.0100**	**0.1469**	**0.6903**	**0.3661**

At inference time, HIDE only queries this frozen generator. For an image x, it obtains a deterministic initial report and additional stochastic reports,

$$y^{(0)} = G_\theta^{\text{greedy}}(x), \qquad y^{(k)} \sim q_\theta(\cdot \mid x), \quad k = 1, \ldots, K.$$

Thus, the sampling distribution q_θ is unchanged throughout the procedure. As described in Sec. 3.1, repeated sampling increases the probability of observing a finding that already has non-zero probability under q_θ, but it cannot create a finding that the frozen generator cannot produce.

HIDE does contain a small post-hoc *corrector-fitting* step. This step estimates the edit-authorization rule; it does not update G_θ or any frozen visual or language representation. For a candidate finding group g and branch b, the corrector uses the fixed features

$$\phi_b(g) = [\text{onehot}(f_g),\, k_b(g),\, m(g)]\,,$$

where f_g is the finding category, $k_b(g)$ is the cross-sample recurrence feature, and $m(g)$ is the prototype margin. The fitted corrector produces

$$q_b(g) = T_{b,\psi_b}(\phi_b(g))\,, \qquad a_b(g) = \mathbf{1}[q_b(g) \geq \tau_b]\,.$$

The parameters ψ_b and thresholds τ_b are estimated once using the designated 0.5% source-domain fitting/calibration split. After this calibration, the corrector is fixed during evaluation.

This fitting operation is a post-hoc decision calibration, rather than training the report generator. It does not change token probabilities, hidden states, future stochastic samples, or the prototype

representations. It only decides whether an already generated candidate should be authorized for an ADD^{+}, ADD^{-}, or DEL action, or whether the system should abstain. Authorized additions copy a source sentence from a stochastic decode, while authorized deletions remove only a matched finding span from the initial draft. Consequently, the corrector cannot invent unrestricted clinical content or freely rewrite the report.

Formally, if ψ collects the parameters of the auxiliary corrector, then

$$\theta^{\mathrm{after}} = \theta^{\mathrm{before}}, \qquad \psi = \mathrm{Fit}(\mathcal{D}_{\mathrm{fit}}; \{\phi_b(g)\}_{b,g}, \text{calibration objective}) .$$

The fitting split is disjoint from the final evaluation split. For cross-dataset transfer to IU-Xray, the source-domain corrector is reused without target-domain fitting, while every report generator remains frozen. Therefore, HIDE is best described as *training-free with respect to the report generator* (or *frozen-generator correction*), while allowing a lightweight post-hoc fitting step for the auxiliary authorization rule. Its gains come from recovering and verifying findings already available in the stochastic outputs of the frozen generator, rather than from retraining that generator.