RECER: REFINING LOCAL TEMPORAL RELATIONS FOR LONGITUDINAL RADIOLOGY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal radiology report generation (LRRG) aims to describe both current findings and their temporal evolution relative to prior examinations. When subsequent finding-specific interpretation relies only on a fixed temporal summary, local cues underrepresented during initial aggregation may remain difficult to incorporate. This motivates refining the interpretation of each prior finding against retained local current–prior comparisons and using the resulting representation to guide both temporal description and historical reuse. To this end, we propose **Re**current Experts for **C**ontrolled **E**vidence **R**euse (**ReCER**), a relation-refinement framework that preserves local current–prior comparison evidence and adaptively refines finding-specific temporal representations, enabling historical reuse to draw on local cues underrepresented in fixed summaries. For each previously reported positive finding, prior-report semantics guide the reading of current–prior visual relations, grounding its reinterpretation in evidence from both examinations. Recurrent relation experts then refine finding-specific temporal representations through state-guided attention, allowing each updated state to guide subsequent evidence aggregation. The refined representations are used to characterize temporal changes, construct temporal content, and derive finding-specific usage scores that control this content's contribution during report generation. ReCER also preserves an independent current-image pathway and accounts for newly emerging findings through a separate branch, so that report generation is not restricted to previously reported positive findings. Experiments on MIMIC-CXR, CheXpert Plus, and IU X-Ray demonstrate the effectiveness of ReCER across longitudinal and current-image-only settings. Our code is available at the anonymous link `https://anonymous.4open.science/r/ReCERcode-6DB5/`.

1 INTRODUCTION

Longitudinal radiology report generation (LRRG) aims to describe current findings and their temporal evolution relative to prior examinations. Prior images provide visual references, while prior reports record earlier observations, allowing the current examination to be interpreted in a temporal context (Hou et al., 2023; Bannur et al., 2023). A previously documented finding, however, cannot be assumed to remain a current fact: it may persist, change, or resolve. The generator must therefore relate historical observations to current visual evidence rather than simply reuse them as additional context. Recent methods incorporate prior examinations through historical feature fusion, cross-examination difference modeling, and disease progression prediction. RECAP jointly models observations and progression (Hou et al., 2023). DDaTR introduces difference-aware representations and adaptive historical fusion, while Diff-RRG uses disease-related local differences to provide progression guidance (Song et al., 2025; Yun et al., 2025). These methods establish the value of explicit comparison. Building on this progress, we ask a more specific question: how can local cross-examination relations be organized around individual findings so that their representations support both temporal description and historical reuse?

Figure 1(a,b) summarizes two complementary emphases: conditioning on historical context and modeling cross-examination changes. They are not mutually exclusive. RECAP gates observation- and progression-related representations (Hou et al., 2023), and DDaTR adaptively incorporates difference-aware historical features (Song et al., 2025). Our concern is not the absence of fusion weights, but the representations on which historical use is based. A progression label characterizes

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.

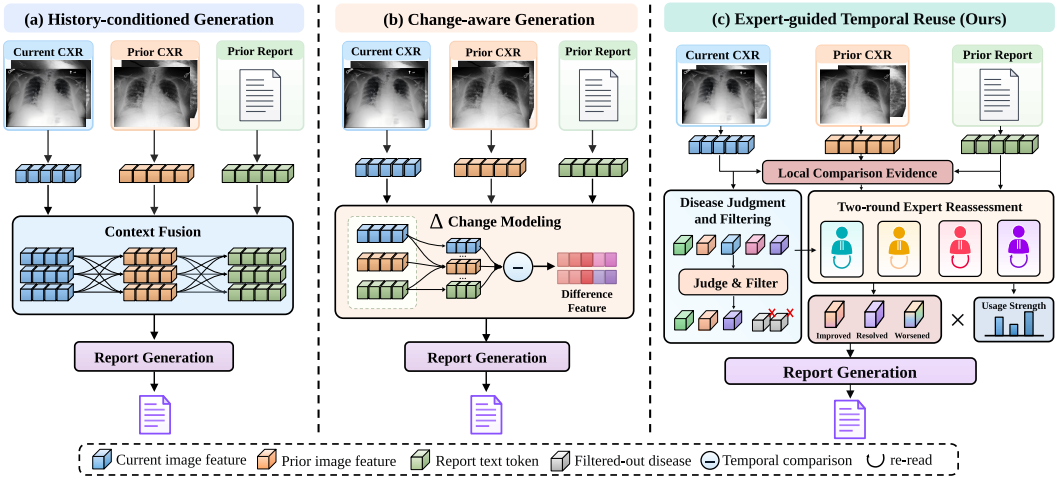


Figure 1: Complementary design emphases for historical information use in LRRG. (a) Conditioning generation on historical context. (b) Guiding generation with cross-examination changes. (c) ReCER refines finding-specific states by revisiting local comparisons and learns how strongly to use the resulting temporal content. NEW uses a separate path, not a fifth recurrent relation expert. CXR denotes chest X-ray.

an inferred change; a fusion weight regulates information flow. Neither alone specifies how the interpretation of an individual finding determines its contribution to generation. ReCER reinterprets prior findings against current evidence. The resulting states guide temporal-content construction and learned usage.

This perspective makes temporal interpretation central to historical reuse. Earlier work already exploits spatial detail: BioViL-T uses local supervision for temporal image classification (Bannur et al., 2023); Diff-RRG aggregates disease-relevant patches for progression guidance (Yun et al., 2025); and PriorRG complements high-level temporal representations with hierarchical visual features (Liu et al., 2026a). The remaining question is whether a developing finding-specific interpretation can continue to access its underlying local comparisons. If subsequent processing receives only a fixed summary, additional transformations can reorganize retained information but cannot revisit relations that are no longer accessible. Cues underrepresented during initial aggregation may therefore remain difficult to incorporate. Retaining position-level relations allows an evolving finding state to redirect evidence aggregation instead of only transforming the initial summary.

ReCER instantiates this principle by constructing shared local relations from current features and their differences from prior visual references. For each previously reported positive finding, prior-report semantics initialize a state and guide relation reading. Recurrent relation experts refine that state while retaining access to the same local evidence, so each update changes how the evidence is aggregated next. Their role is to develop a finding's temporal interpretation, not merely deepen transformations of a fixed representation. The refined states characterize change, construct temporal content, and yield finding-specific usage scores. Learned through report generation, these scores modulate the corresponding temporal records during decoding. Thus, the interpretation formed from current–prior evidence also informs how that content contributes to the report.

To support this process, ReCER employs a shared Visual Representation Extraction (VRE) front end that retains position-level visual features from each examination. These features support current-finding recognition and the construction of local current–prior relations that remain accessible during finding-specific refinement. VRE thus provides a common visual basis for temporal interpretation while preserving current-image evidence independently of historical inputs.

Our main contributions are summarized as follows:

- **A finding-level perspective on historical reuse.** Historical reuse in LRRG is framed around reinterpreting prior findings against current evidence, connecting their temporal evolution to their contribution to the current report.

- **Expert-driven relation refinement for controlled reuse.** We propose ReCER with recurrent relation experts that refine finding-specific temporal representations through state-guided rereading of retained local current–prior evidence. These representations support temporal-content construction and controlled historical reuse during report generation.
- **A shared visual foundation for finding recognition and temporal comparison.** VRE combines local adaptation, within-image contralateral context, and frozen semantic refinement, providing spatially resolved features for current-finding recognition, local current–prior comparison, and finding-specific relation refinement.
- **Promising report-generation performance.** Experiments on MIMIC-CXR, CheXpert Plus, and IU X-Ray show promising report-level language quality and clinical label agreement across longitudinal and current-image-only settings.

2 Related Work

2.1 Historical Context Utilization

Learned memory and patient history provide different forms of context. R2Gen and R2GenCMN use relational or shared cross-modal memory for image–text modeling (Chen et al., 2020; 2021). Longitudinal systems instead condition on the patient's earlier examinations: Prefilling fuses prior images and reports (Zhu et al., 2023), CXRMate incorporates earlier reports and multi-image studies (Nicolson et al., 2024), and HERGen aggregates imaging histories through group-causal attention (Wang et al., 2024). MLRG learns multi-view longitudinal representations through contrastive objectives (Liu et al., 2025a).

STREAM combines spatiotemporal modeling with region-level semantic retrieval (Yang et al., 2025). PriorRG couples prior-guided contrastive pretraining with coarse-to-fine decoding (Liu et al., 2026a). These approaches improve the acquisition and integration of context. Historical conditioning alone, however, does not specify how the interpretation of each prior finding should determine its contribution after comparison with current evidence. ReCER links each prior-positive finding's contribution to its visually refined state, enabling selective historical reuse alongside independent current-image sources.

2.2 Temporal Comparison and Refinement

BioViL-T learns temporal vision–language representations, and controllable reporting compares anatomical regions across examinations (Bannur et al., 2023; Dalla Serra et al., 2023). RECAP models observations and progression with gated decoding (Hou et al., 2023); HC-LLM constrains shared and time-specific information (Liu et al., 2025b); DDaTR fuses difference-aware residual features (Song et al., 2025). Local comparison and learned weighting are therefore established ingredients. Diff-RRG uses disease-related patch differences as progression guidance (Yun et al., 2025), Libra aligns temporal features through a connector (Zhang et al., 2025), and CoCa-CXR restricts cross-image attention spatially (Chen et al., 2025). MARE models disease-evolution relations (Gao et al., 2026), while BiOTPrompt uses bidirectional transport asymmetry to identify emerging and resolved regions for dynamic prompts (Liu et al., 2026b).

The object and timing of refinement distinguish closely related designs. TIM iteratively corrects prior and current reports through mutual feedback (Dong et al., 2026). STRIVE reconciles diagnostic, attribute, and temporal agent outputs before generation, then validates the resulting report (Maeng et al., 2026). These mechanisms can coexist with historical conditioning, as indicated by the overlapping emphases in Figure 1. ReCER refines representations of eligible prior-positive findings through state-guided reading of local current–prior relations before decoding. The refined states determine both the content of the corresponding temporal records and their usage during generation, linking the reinterpretation of prior findings to controlled historical reuse.

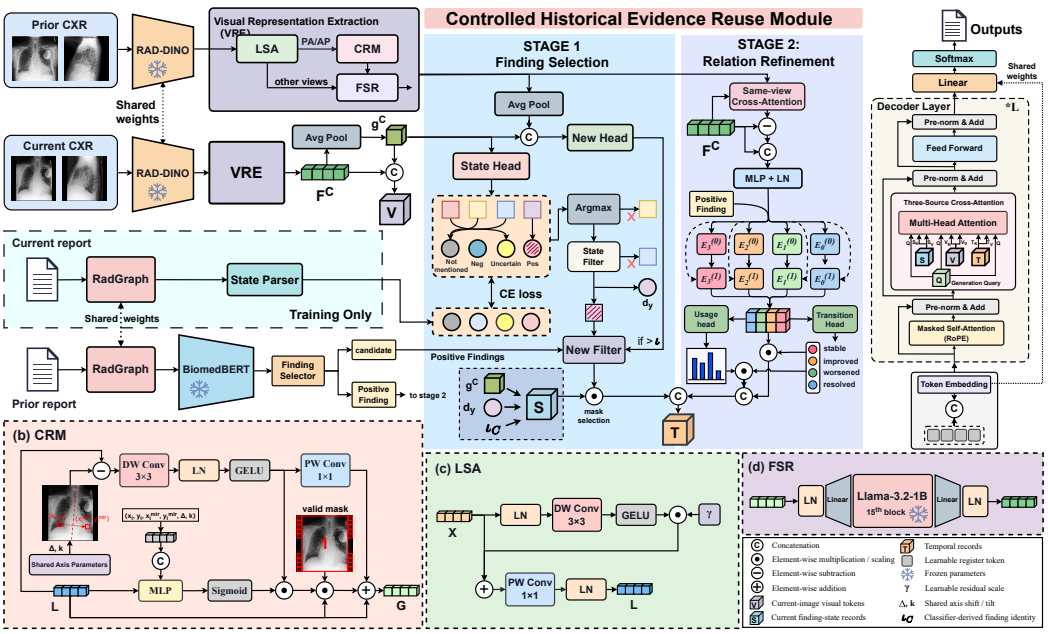


Figure 2: Overview of ReCER. Shared VRE features support current-finding recognition and local current–prior comparison. Recurrent experts refine prior-positive finding states to construct OLD records and learn their usage; a separate NEW path retrieves eligible current-positive records. The decoder combines current finding records (S), current visual features (V), and temporal records (T). Current reports provide supervision only.

3 Method

3.1 Task Formulation and Overview

Let I_C and r_C denote the current images and target report. The pair (I_P, r_P) comprises the images and report from the most recent eligible study of the same patient and is shared by all findings. ReCER models

$$p_\theta(r_C \mid I_C, I_P, r_P) = \prod_{\ell=1}^{L} p_\theta(w_\ell \mid w_{<\ell}, S, V, T). \tag{1}$$

Here L is report length; S, V, and T contain current finding records, current visual features, and temporal records. Without an eligible prior comparison, $T = \varnothing$ and generation uses S, V. Figure 2 summarizes the architecture. All source and state vectors have width $d = 512$; N_0 and LN denote non-affine and affine layer normalization, respectively. MLP denotes a multilayer perceptron. Target reports provide supervision only. The retained local relations R remain fixed during refinement; the finding-specific states that read them are updated.

3.2 Local Visual Feature Extraction

Our shared VRE front end combines three modules after a frozen RAD-DINO encoder (Pérez-García et al., 2025). The Local Spatial Adapter (LSA) adds a depthwise-convolution residual on the patch grid, then projects 768 channels to 512. Contralateral Reference Modulation (CRM) samples a within-image reference for each patch at a position obtained by reflecting the patch center across a learned, shared axis and applying a local offset. A gated local-minus-reference transformation updates each frontal-image feature; the local feature is preserved when its reference is invalid, and other views bypass CRM.

Frozen Semantic Refinement (FSR) follows the frozen-block adaptation principle of Tang et al. (2025). Trainable adapters connect visual features to the frozen fifteenth Llama-3.2-1B block for bidirectional transformation, preserving gradient flow through it. Within FSR, language causal masking and

rotary position embeddings are disabled; all valid spatial positions remain available for comparison. The three operations serve complementary roles: local adaptation preserves neighborhood structure, contralateral modulation introduces a within-image reference, and semantic refinement transforms features without pooling away their spatial positions. Both examinations pass through the same front end, so cross-study matching operates in a shared feature space. The outputs F_C, F_P and their valid-patch means g_C, g_P are shared downstream. Appendix A.2 specifies the operators; Appendix A.2.1 gives CRM's geometry and implementation algorithm.

3.3 Controlled Historical Evidence Reuse

3.3.1 Finding Selection

A linear State Head maps g_C to four logits per anatomically scoped finding key c: positive, negative, uncertain, and not mentioned. Let $\widehat{y}_c$ be their argmax, $w_{c,y}$ the classifier weight for key c and state y, and d_y a shared state-type embedding. For keys predicted as positive, negative, or uncertain,

$$\iota_c = N_0(w_{c,\text{positive}} - w_{c,\text{negative}}), \qquad s_c = N_0(g_C + \iota_c + d_{\widehat{y}_c}). \tag{2}$$

Thus $S = \{s_c : \widehat{y}_c \neq \text{not mentioned}\}$ retains negative and uncertain records. Gradients reach classifier weights through ι_c.

RadGraph-based parsing (Jain et al., 2021) supplies prior assertions, anatomical scope, and text spans. OLD requires a valid prior-positive key, usable text, and a compatible same-view comparison; current positivity is unnecessary, allowing resolved findings. For NEW, candidate keys must have a valid prior state of negative or not mentioned and complete anatomical scope; uncertain or conflicting prior assertions are excluded. A sigmoid-activated linear head applied to $[g_C; g_P]$ produces p_c^{new}. Emission requires all three conditions: candidate eligibility, $\widehat{y}_c = \text{positive}$, and $p_c^{\text{new}} \geq \tau$, with a shared threshold $\tau = 0.8$ for prior-negative (NEG) and prior-not-mentioned (NM) candidates. For emitted keys,

$$t_c^{\text{new}} = s_c + e_{\text{new}}, \qquad g_c^{\text{new}} = p_c^{\text{new}}. \tag{3}$$

The existing s_c is retrieved by key, and e_{new} is a shared type embedding. NEW bypasses OLD refinement because its content is drawn from the current positive record rather than the reinterpretation of a prior-positive finding. OLD and NEW are disjoint by prior-state eligibility, while both can coexist with the independent current-image sources. Prior nonmention does not imply absence; positive NEW supervision requires explicit, reliably reference-bound emergence evidence (Appendix A.7). The threshold admits NEW records without imposing hard constraints on decoder wording. The decoder uses S, V, and available T, so admitting a NEW record does not ensure a correct temporal statement.

3.3.2 Relation Refinement

Retaining local comparisons. For current patch i, Ω_i contains all compatible prior patches from posteroanterior (PA)–PA, anteroposterior (AP)–AP, or compatible lateral–lateral image pairs. With bias-free projections $W_Q^X, W_K^X \in \mathbb{R}^{d\times d}$, soft matching and an MLP with a Gaussian error linear unit (GELU) activation construct

$$\widehat{f}_i^{P\to C} = \sum_{u\in\Omega_i} a_{iu} f_u^P, \qquad R_i = \text{LN}_{\text{rel}}\Big(\text{MLP}_{\text{rel}}([f_i^C; f_i^C - \widehat{f}_i^{P\to C}])\Big). \tag{4}$$

The superscript X identifies cross-examination matching. Here $a_{iu} = \text{softmax}_{u\in\Omega_i}((f_i^C W_Q^X)(f_u^P W_K^X)^\top/\sqrt{d})$. Relations for $\mathcal{I}_R = \{i : \Omega_i \neq \varnothing\}$ are computed once and shared across findings and updates; unmatched current patches remain in V. Each retained relation is indexed by a current patch and represents its comparison with a softly aggregated prior reference. Subsequent finding-specific reads therefore revisit these position-level relations, not the full set of original current–prior patch pairs.

State-guided reading. Frozen BiomedBERT (Gu et al., 2021) encodes the prior report. Selected tokens are projected, normalized, and averaged over unique original positions to initialize h_c^0 (Appendix A.3). All four branches start at $z_{c,j}^{(0)} = h_c^0$, without a visual pre-read. For branch $j \in \{0,1,2,3\}$ and update $r \in \{1,2\}$,

$$K_i^E = N_0(R_i)W_K^E, \qquad Q_{c,j}^{(r)} = N_0(z_{c,j}^{(r-1)} + \rho_j)W_Q^E. \tag{5}$$

$$\alpha_{c,j,i}^{(r)} = \mathrm{softmax}_{i \in \mathcal{I}_R}\left(\frac{Q_{c,j}^{(r)}(K_i^E)^\top}{\sqrt{d_q}}\right), \qquad b_{c,j}^{(r)} = \sum_{i \in \mathcal{I}_R} \alpha_{c,j,i}^{(r)} R_i. \tag{6}$$

$$x_{c,j}^{(r)} = N_0\Big(b_{c,j}^{(r)} + N_0(z_{c,j}^{(r-1)})W_Z\Big), \qquad z_{c,j}^{(r)} = b_{c,j}^{(r)} + \mathcal{E}_j(x_{c,j}^{(r)}). \tag{7}$$

The shared query/key projections have width $d_q = 64$; $\rho_j \in \mathbb{R}^d$ is a branch identity. GELU is also used in the expert MLPs. The state projection $W_Z \in \mathbb{R}^{d \times d}$ is shared; each branch's MLP $\mathcal{E}_j : d \to 2048 \to d$ is reused across updates. Thus an updated state redirects the next read of the same R. All branches execute without gradient detachment. The retained relations are shared, but the queries depend on each finding state; different findings can therefore aggregate different positions from the same image pair. Across updates, the experts reuse parameters and read the same relation bank.

Temporal content and usage. Concatenated final states feed a hierarchical transition head that predicts persistence versus resolution and, conditional on persistence, stable, improved, or worsened states. Multiplying the persistence probability by each conditional probability yields the probabilities of stable, improved, and worsened states; the resolution probability supplies the resolved entry of the ordered probability vector π_c (Appendix A.4). These probabilities are used to construct one OLD record and its usage score:

$$\begin{aligned} t_c^{\mathrm{old}} &= N_0\left(\iota_c + \mathrm{LN}_{\mathrm{fuse}}\left(\sum_{j=0}^{3} \pi_{c,j} z_{c,j}^{(2)}\right)\right), \\ g_c^{\mathrm{old}} &= \sum_{j=0}^{3} \pi_{c,j}\, \sigma\Big(N_0(z_{c,j}^{(2)})w_g + b_g\Big). \end{aligned} \tag{8}$$

Here, $w_g \in \mathbb{R}^d$ and $b_g \in \mathbb{R}$ define a shared usage head, and σ is the sigmoid function. OLD usage is learned through report generation to modulate contribution; it is not a calibrated probability that a finding or its temporal interpretation is correct. Content and log-usage remain separate; valid OLD records are not thresholded. OLD and NEW form T; prior text never enters the decoder directly.

3.4 Source-Aware Report Generation

A four-layer, pre-normalized variant of the Transformer (Vaswani et al., 2017) with eight heads reads S, V, T, where V contains all valid current patches and their mean. Four register states follow the causal visibility rules in Appendix A.5. For source a, entry i, and head b, let $q_b, k_{a,i,b}, v_{a,i,b} \in \mathbb{R}^{64}$ be projected queries, keys, and values. Set $g_i = 1$ for current sources; temporal records carry learned usage:

$$\ell_{a,i,b} = \frac{q_b k_{a,i,b}^\top}{\sqrt{d_h}} + \log g_i, \qquad d_h = 64. \tag{9}$$

The within-source context is $C_{a,b} = \sum_i \mathrm{softmax}_i(\ell_{a,i,b}) v_{a,i,b}$. For the nonempty-source set $\mathcal{A}$, source competition uses the log-sum-exp operator LSE:

$$m_{a,b} = \mathrm{LSE}_i(\ell_{a,i,b}) - \log n_a + \delta_a, \qquad \beta_{a,b} = \mathrm{softmax}_{a \in \mathcal{A}}(m_{a,b}). \tag{10}$$

Here, n_a is a source-normalization factor: $n_S = |S|$, $n_V = |V|$, and $n_T = 1$; δ_a is a learned bias shared across heads. Each OLD or NEW record remains a separate key–value entry with its own log-usage-adjusted attention logit. Setting $n_T = 1$ preserves summed, usage-weighted temporal contributions in source competition, making temporal-source strength depend on record count; no pre-attention pooling occurs. The cross-attention output is

$$Y = \sum_{a \in \{S,V,T\}} \mathrm{cat}_{b=1}^{8}(\beta_{a,b} C_{a,b})\, W_a^O. \tag{11}$$

Here $W_a^O \in \mathbb{R}^{d \times d}$. Reusing adjusted logits makes usage affect both within-source allocation and source competition without multiplying values again. Appendix A.5 derives the effective weights and their record-count dependence.

3.5 Training Objective

Report generation and auxiliary tasks are jointly optimized:

$$\mathcal{L} = \mathcal{L}_{\text{report}} + \mathcal{L}_{\text{state4}} + \mathcal{L}_{\text{transition4}} + \mathcal{L}_{\text{new}}. \tag{12}$$

The terms are report negative log-likelihood, four-state cross-entropy, transition cross-entropy, and NEW binary cross-entropy. Four-state targets map report assertions to shared anatomically scoped finding keys. Temporal targets require comparisons reliably linked to the selected prior; negative-to-positive or not-mentioned-to-positive changes alone do not establish emergence. Conflicting assertions, unresolved scope, and ambiguous temporal references mask affected targets. Missing transition targets suppress only their loss, retaining otherwise eligible OLD records. NEW supervision is independent of runtime candidate filtering and emission; current reports supervise training only. Each loss averages over valid supervision units; empty tasks contribute zero. Appendix A.6 specifies optimization; Appendix A.7 defines supervision domains.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation. MIMIC-CXR (Johnson et al., 2019a) and CheXpert Plus (Chambon et al., 2024) provide longitudinal settings; IU X-Ray (Demner-Fushman et al., 2016) provides current-image-only evaluation. Pairing respects the supplied partitions and uses the latest eligible prior; cases without history remain included. Appendix A.1 provides the dataset release statistics. We report bilingual evaluation understudy (BLEU) at order four, metric for evaluation of translation with explicit ordering (METEOR), recall-oriented understudy for gisting evaluation (ROUGE) using its longest-common-subsequence variant (ROUGE-L), and clinical efficacy (CE) precision (CE-P), recall (CE-R), and F1 (CE-F1) (Papineni et al., 2002; Banerjee and Lavie, 2005; Lin, 2004; Smit et al., 2020). ReCER uses micro-averaged CE metrics. Appendix A.8 describes the evaluation metrics.

Implementation. The default configuration uses four experts, two updates, and four decoder layers. The specified AdamW (Loshchilov and Hutter, 2019) training configuration uses 30 epochs, a batch size of 16, a peak learning rate of 10^{-4}, 5% warmup, cosine decay, and bfloat16 (BF16) precision (Appendix A.6). Frozen encoder weights are excluded from optimization while gradients through the frozen FSR block remain available to its adapters.

4.2 Comparison Across Three Datasets

Table 1 summarizes ReCER and literature results. ReCER achieves BLEU-4/METEOR of 0.178/0.190, 0.134/0.158, and 0.211/0.234 on MIMIC-CXR, CheXpert Plus, and IU X-Ray, respectively, with micro-F1 of 0.523, 0.355, and 0.576. These results support reporting with and without history while retaining independent current-image evidence.

On MIMIC-CXR, ReCER's micro-F1 is 0.523, versus reported values of 0.511 (PriorRG), 0.505 (MLRG), and 0.484 (R2GenCSR). Clinical label agreement is favorable, subject to protocol differences (Appendix F). Compared with PriorRG, precision is higher (0.576 versus 0.541), but recall lower (0.479 versus 0.485). Table 1 cannot isolate architectural contributions; subsequent ablations evaluate components within our setting.

4.3 Ablation Study on MIMIC-CXR

The full model achieves the highest CE-F1 and the highest scores on all three language metrics among variants defined in Appendix D (Table 2), with CE-F1 gains of 0.008 over expert replacement and 0.012 over joint visual replacement. Expert replacement retains local relations, current sources, and NEW, fills all four posterior slots with one ordinary read, and removes branch-specific updates. Higher precision (0.576 versus 0.557) at equal recall supports the overall refinement design.

Visual replacements affect current reporting and temporal comparison jointly; effects are nonadditive. FSR replacement removes the frozen block and adapters without isolating language-model pretraining. After removing the learned head, threshold, and NEW loss, "w/o NEW head" still emits eligible

Table 1: Comparison of report generation performance on MIMIC-CXR, CheXpert Plus, and IU X-Ray.

Method	Year	History	BLEU-4	METEOR	ROUGE-L	CE-P	CE-R	CE-F1
MIMIC-CXR								
R2Gen	2020		0.103	0.142	0.277	0.333	0.273	0.276
R2GenCMN	2021		0.106	0.142	0.278	0.334	0.275	0.278
CXRMate	2024	✓	0.079	–	0.262	0.438	0.349	0.357
STREAM‡	2025	✓	0.139	0.172	0.297	0.515	0.447	0.478
MLRG	2025a	✓	0.158	0.176	0.320	0.549	0.468	0.505
REVA-PO	2026		0.143	0.170	0.293	0.569	0.455	0.506
R2GenCSR	2026b		0.136	0.167	0.291	0.551	0.432	0.484
PriorRG	2026a	✓	0.175	0.189	0.324	0.541	**0.485**	0.511
ReCER (ours)	2026	✓	**0.178**	**0.190**	**0.329**	**0.576**	0.479	**0.523**
CheXpert Plus								
LR2Gen*	2023	✓	0.074	0.128	0.235	0.217	0.308	0.205
MambaXray-VL-B*	2025		0.105	0.149	0.267	0.333	0.264	0.273
R2GenCSR*	2026b		0.100	0.146	0.265	0.315	0.247	0.259
ReCER (ours)	2026	✓	**0.134**	**0.158**	**0.323**	**0.388**	**0.327**	**0.355**
IU X-Ray								
DDaTR†	2025		0.103	0.160	0.307	0.304	0.291	0.262
STREAM	2025		0.188	0.215	0.387	–	–	–
MambaXray-VL-B*	2025		0.179	0.215	0.388	–	–	–
AM-MRG	2026a		0.192	0.225	0.384	–	–	–
R2GenCSR	2026b		0.206	0.215	**0.401**	–	–	–
ReCER (ours)	2026		**0.211**	**0.234**	0.397	**0.557**	**0.598**	**0.576**

*From CXPMRG-Bench (Wang et al., 2025): LR2Gen and R2GenCSR are benchmark reproductions; MambaXray-VL-B is the benchmark authors' model. †DDaTR is trained on MIMIC-CXR and tested on IU X-Ray without IU training or fine-tuning (Song et al., 2025). ‡STREAM with prior reports.

Table 2: Ablation study of ReCER on MIMIC-CXR. Column bests are bold, including ties.

Configuration	BLEU-4	METEOR	ROUGE-L	CE-P	CE-R	CE-F1
Full	**0.178**	**0.190**	**0.329**	**0.576**	**0.479**	**0.523**
w/o experts	0.160	0.177	0.315	0.557	**0.479**	0.515
w/o CRM	0.167	0.182	0.318	0.571	0.465	0.513
w/o FSR	0.175	0.172	0.326	0.569	0.473	0.517
w/o LSA	0.173	0.184	0.321	0.575	0.473	0.519
w/o CRM+FSR+LSA	0.163	0.174	0.314	0.563	0.467	0.511
w/o NEW head	0.169	0.181	0.323	0.561	0.476	0.515
w/o registers	0.176	0.187	0.327	0.574	0.478	0.522

current-positive NEW records with unit usage. The CE-F1 drop to 0.515 supports the combined NEW design; register removal yields 0.522. Other-dataset results appear in Appendix B.

4.4 Parameter Analysis on MIMIC-CXR

In the one-factor analysis (Figure 3; full values in Appendix B.1), the default four experts, two updates, and four decoder layers match or exceed every tested alternative on each reported metric, with CE-F1 of 0.523.

Five experts match four on METEOR but trail on other metrics; three and six also underperform. More branches thus offer no consistent benefit with four-slot mixing. Two updates improve all language metrics and CE-F1 over one (0.523 versus 0.519), but three reduce CE-F1 to 0.516. This favors a second finding-guided read using shared parameters. Neither further refinement nor a fifth decoder layer helps.

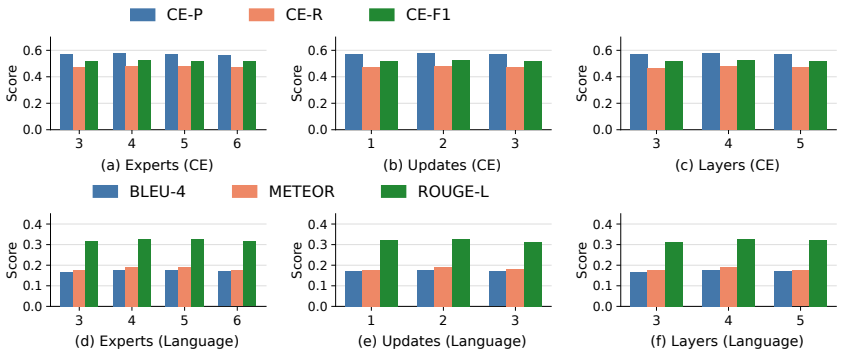


Figure 3: MIMIC-CXR parameter sensitivity results. Columns vary expert count, update count, and decoder depth. Rows show CE and language metrics; full values appear in Appendix B.1.

Parameter count. Parameter profiles report 310.28 million (M) total/53.99M trainable parameters for the shared MIMIC-CXR/CheXpert Plus configuration and 201.03M total/42.10M trainable parameters for the IU X-Ray configuration, which omits BiomedBERT. Appendix C gives the module-level counts and report-generation parameter comparisons.

4.5 Qualitative Analysis

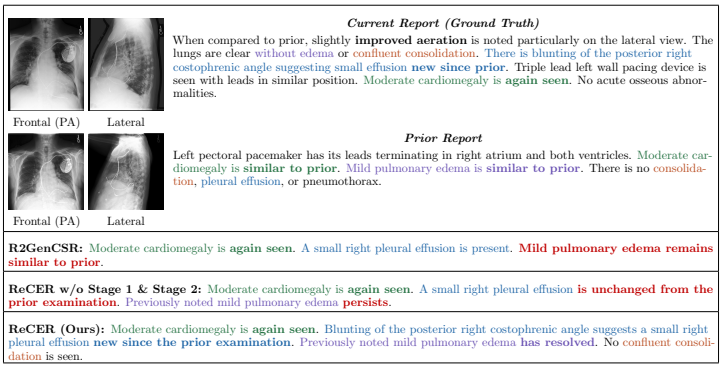


Figure 4: Qualitative comparison of longitudinal report generation on MIMIC-CXR. Matching colors indicate corresponding findings; bold red text marks descriptions inconsistent with the displayed reference report.

In Figure 4, the current report suggests a new small right pleural effusion, excludes edema, and again describes cardiomegaly. The baseline carries edema forward; the variant without finding selection and relation refinement also calls the effusion unchanged. ReCER preserves uncertainty and distinguishes possible emergence from edema resolution while retaining persistent findings. This illustrates finding-specific historical reuse through reinterpretation against current evidence. Further examples and temporal ambiguities appear in Appendix E. The qualitative "w/o Stage 1 & Stage 2" variant should not be equated with Table 2's "w/o experts" replacement, whose retained components are defined in Appendix D.1.

5 Conclusion

ReCER retains local current–prior relations and uses updated finding states to guide evidence aggregation, temporal-content construction, and finding-specific usage. VRE supplies shared local representations, while independent current-image sources and a separate NEW path preserve coverage beyond prior abnormalities. This framework connects finding-specific reinterpretation with controlled historical reuse for LRRG.

AI USE STATEMENT

GPT-6 Astra assisted with manuscript preparation, including language editing for clarity, concision, and readability. The authors reviewed all AI-assisted revisions to preserve the intended scientific meaning and take full responsibility for the final manuscript.

ETHICS STATEMENT

This study uses public chest X-ray datasets for research on radiology report generation and is not intended for autonomous clinical decision-making. Dataset access and use are subject to the conditions specified by the dataset providers.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A specify the computation, input eligibility, visual and decoder interfaces, supervision masks, and optimization configuration. The CRM convention and algorithm are given in Appendix A.2.1; component replacements and sensitivity interfaces are defined in Appendix D. The accompanying source package includes the LaTeX manuscript, tables, and figures. The code for each module is available at the anonymous link `https://anonymous.4open.science/r/ReCERcode-6DB5/`.

REFERENCES

Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72, 2005.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing. *CVPR*, 15016–15027, 2023. DOI: 10.1109/CVPR52729.2023.01442.

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:2405.19538*, 2024.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating Radiology Reports via Memory-driven Transformer. *EMNLP*, 1439–1449, 2020. DOI: 10.18653/v1/2020.emnlp-main.112.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal Memory Networks for Radiology Report Generation. *ACL-IJCNLP*, 5904–5914, 2021. DOI: 10.18653/v1/2021.acl-long.459.

Yixiong Chen, Shawn Xu, Andrew Sellergren, Yossi Matias, Avinatan Hassidim, Shravya Shetty, Daniel Golden, Alan L. Yuille, and Lin Yang. CoCa-CXR: Contrastive Captioners Learn Strong Temporal Structures for Chest X-Ray Vision-Language Understanding. *MICCAI*, 78–88, 2025. DOI: 10.1007/978-3-032-04978-0_8.

Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison O'Neil. Controllable Chest X-Ray Report Generation from Longitudinal Representations. *Findings of EMNLP*, 4891–4904, 2023. DOI: 10.18653/v1/2023.findings-emnlp.325.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. DOI: 10.1093/jamia/ocv080.

Yiheng Dong, Yi Lin, Shilong Huang, Xiyan Yang, and Xin Yang. TIM: Temporal Decoupling with Iterative Mutual-Refinement Model for Longitudinal Radiology Report Generation. *CVPR*, 6951–6961, 2026.

Qingqing Gao, Tengfei Liu, Xiaoyan Li, Xiaodan Zhang, Zhongfan Sun, Boyue Wang, Baocai Yin, and Zhaohui Liu. MARE: Multimodal Analogical Reasoning for Disease Evolution-Aware Radiology Report Generation. *AAAI*, 40(25):21180–21188, 2026. DOI: 10.1609/aaai.v40i25.39262.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021 (online publication). DOI: 10.1145/3458754.

Li Guo, Anas M. Tahir, and Z. Jane Wang. REVA-PO: Stabilizing Reinforcement Learning for Chest X-ray Report Generation. *ECCV*, 2026. arXiv:2607.10147v2.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. RECAP: Towards Precise Radiology Report Generation via Dynamic Disease Progression Reasoning. *Findings of EMNLP*, 2134–2147, 2023. DOI: 10.18653/v1/2023.findings-emnlp.140.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *NeurIPS Datasets and Benchmarks*, 2021.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019a. DOI: 10.1038/s41597-019-0322-0.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv:1901.07042*, 2019b.

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, 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. *CVPR*, 10348–10359, 2025a. DOI: 10.1109/CVPR52734.2025.00968.

Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. HC-LLM: Historical-Constrained Large Language Models for Radiology Report Generation. *AAAI*, 39(6):5595–5603, 2025b. DOI: 10.1609/aaai.v39i6.32596.

Kang Liu, Zhuoqi Ma, Zikang Fang, Yunan Li, Kun Xie, and Qiguang Miao. PriorRG: Prior-Guided Contrastive Pre-training and Coarse-to-Fine Decoding for Chest X-ray Report Generation. *AAAI*, 40(9):7206–7214, 2026a. DOI: 10.1609/aaai.v40i9.37657. Extended version with supplementary Table 7: `https://arxiv.org/abs/2508.05353v2`.

Tengfei Liu, Yijian Fan, Boyue Wang, Yongli Hu, Mingjie Li, Jinghua Li, Junbin Gao, Xiaojun Chang, Zhihui Li, and Baocai Yin. BiOTPrompt: Bidirectional Optimal Transport Guided Prompting for Disease Evolution-aware Radiology Report Generation. *CVPR*, 13755–13765, 2026b.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. *ICLR*, 2019. arXiv:1711.05101.

Junyeong Maeng, Eunsong Kang, and Heung-Il Suk. STRIVE: Multi-Agent Structured Temporal Reasoning with Integrated Verification for Longitudinal Radiology Report Generation. *arXiv:2608.24237*, 2026.

Meta. Llama 3.2 1B model card. 2024. `https://huggingface.co/meta-llama/Llama-3.2-1B`. Accessed September 25, 2026.

Aaron Nicolson, Jason Dowling, Douglas Anderson, and Bevan Koopman. Longitudinal Data and a Semantic Similarity Reward for Chest X-Ray Report Generation. *Informatics in Medicine Unlocked*, 50:101585, 2024. DOI: 10.1016/j.imu.2024.101585.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. *ACL*, 311–318, 2002. DOI: 10.3115/1073083.1073135.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7:119–130, 2025. DOI: 10.1038/s42256-024-00965-w.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew P. Lungren. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. *EMNLP*, 1500–1519, 2020. DOI: 10.18653/v1/2020.emnlp-main.117.

Shanshan Song, Hui Tang, Honglong Yang, and Xiaomeng Li. DDaTR: Dynamic Difference-aware Temporal Residual Network for Longitudinal Radiology Report Generation. *IEEE Transactions on Medical Imaging*, 44(12):5345–5357, 2025. DOI: 10.1109/TMI.2025.3591364.

Stanford Center for Artificial Intelligence in Medicine and Imaging. CheXpert Plus: dataset documentation. n.d. `https://aimi.stanford.edu/datasets/chexpert-plus`. Accessed September 25, 2026. Dataset DOI: 10.71718/6nvz-pm34.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. DOI: 10.1016/j.neucom.2023.127063.

Fenghe Tang, Wenxin Ma, Zhiyang He, Xiaodong Tao, Zihang Jiang, and S. Kevin Zhou. Pre-Trained LLM is a Semantic-Aware and Generalizable Segmentation Booster. *MICCAI*, 402–412, 2025. DOI: 10.1007/978-3-032-05127-1_39.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *NeurIPS*, 30, 2017.

Fuying Wang, Shenghui Du, and Lequan Yu. HERGen: Elevating Radiology Report Generation with Longitudinal Data. *ECCV*, 183–200, 2024. DOI: 10.1007/978-3-031-73001-6_11.

Xiao Wang, Fuling Wang, Yuehang Li, Qingchuan Ma, Shiao Wang, Bo Jiang, and Jin Tang. CXPMRG-Bench: Pre-training and Benchmarking for X-ray Medical Report Generation on CheXpert Plus Dataset. *CVPR*, 5123–5133, 2025.

Xiao Wang, Fuling Wang, Haowen Wang, Bo Jiang, Chuanfu Li, Yaowei Wang, Yonghong Tian, and Jin Tang. Activating Associative Disease-Aware Vision Token Memory for LLM-Based X-ray Report Generation. *IEEE Transactions on Medical Imaging*, 45(2):583–595, 2026a. DOI: 10.1109/TMI.2025.3603416.

Xiao Wang, Yuehang Li, Fuling Wang, Shiao Wang, Chuanfu Li, and Bo Jiang. R2GenCSR: Mining Contextual and Residual Information for LLMs-based Radiology Report Generation. *IEEE Journal of Biomedical and Health Informatics*, 30(9):7667–7680, 2026b. DOI: 10.1109/JBHI.2026.3669539. arXiv:2408.09743v2.

Yan Yang, Xiaoxing You, Ke Zhang, Zhenqi Fu, Xianyun Wang, Jiajun Ding, Jiamei Sun, Zhou Yu, Qingming Huang, Weidong Han, and Jun Yu. Spatio-Temporal and Retrieval-Augmented Modeling for Chest X-Ray Report Generation. *IEEE Transactions on Medical Imaging*, 44(7):2892–2905, 2025. DOI: 10.1109/TMI.2025.3554498.

Hannah Yun, Junyeong Maeng, Eunsong Kang, and Heung-Il Suk. Diff-RRG: Longitudinal Disease-Wise Patch Difference as Guidance for LLM-Based Radiology Report Generation. *MICCAI*, 152–161, 2025. DOI: 10.1007/978-3-032-04981-0_15.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. Libra: Leveraging Temporal Images for Biomedical Radiology Analysis. *Findings of ACL*, 17275–17303, 2025. DOI: 10.18653/v1/2025.findings-acl.888.

Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M. Summers, and Zhiyong Lu. Utilizing Longitudinal Chest X-Rays and Reports to Pre-Fill Radiology Reports. *MICCAI*, 189–198, 2023. DOI: 10.1007/978-3-031-43904-9_19.

A Implementation and Evaluation Details

A.1 Datasets, partitions, and study pairing

Table 3 distinguishes dataset scale, counting unit, and split source. MIMIC-CXR-JPG provides an official patient-disjoint split with 222,758 training studies, 1,808 validation studies, and 3,269 test studies (Johnson et al., 2019b). Its 227,835 reports include eight without text usable by the original labeling pipeline; the smaller label-file total is not a different study split. The corresponding image counts include all released views.

The Stanford Center for Artificial Intelligence in Medicine and Imaging reports 223,462 image–report pairs across 187,711 studies and 64,725 patients for CheXpert Plus (Stanford AIMI, n.d.). Its dataset-scale counts must be distinguished from the filtered Findings benchmark in CXPMRG-Bench, which reports 40,463, 5,780, and 11,562 training, validation, and test samples (Wang et al., 2025). For IU X-Ray, the source-scale figures reported in that benchmark are 7,470 images and 3,955 reports. The paired-view benchmark documented by PriorRG contains 2,069, 296, and 590 reports, with two images per report (Liu et al., 2026a). The latter two benchmark partitions are published reference protocols.

Table 3: Public dataset statistics and documented partitions.

Dataset / source scale		Images		Reports
MIMIC-CXR		377,110		227,835
CheXpert Plus		223,462		187,711
IU X-Ray		7,470		3,955
Dataset / split source	**Counting unit**	**Train**	**Validation**	**Test**
MIMIC-CXR-JPG	Patients	64,586	500	293
	Studies / reports	222,758	1,808	3,269
	Images	368,960	2,991	5,159
CheXpert Plus	Findings samples	40,463	5,780	11,562
IU X-Ray	Studies / reports	2,069	296	590
	Images	4,138	592	1,180

We retain the supplied partition assignments and link prior studies only within a patient's partition. Every finding in a current study refers to the same most recent prior that passes the study-level image/report availability checks; view and finding eligibility are evaluated afterward. Missing text spans or views do not trigger selection of an older study. Valid current images remain available even when they lack a compatible prior view. PA–PA, AP–AP, and compatible lateral–lateral comparisons are permitted; PA–AP and frontal–lateral comparisons are excluded. Additional acquisition restrictions may remove, but not introduce, a pairing. Absence of an eligible comparison invokes current-image-only generation rather than excluding the sample.

The public counts in Table 3 describe their stated release or benchmark. IU evaluation here uses current images only and is not presented as evidence of zero-shot temporal transfer.

A.2 VRE modules and CRM implementation

VRE applies local adaptation, within-image reference modulation, and semantic refinement in sequence. Its output remains a set of position-level features, shared by current-finding recognition and longitudinal relation construction. The following module descriptions explain how that interface is maintained; the CRM algorithm appears in the same subsection.

Images are resized with aspect ratio preserved and symmetrically padded to 518×518. RAD-DINO supplies a 37×37 grid of 768-dimensional patch features. Pure-padding patches are invalid; partially covered patches remain valid. Invalid positions are masked before spatial mixing and again after affine normalization where needed. Each image is processed independently with shared weights.

LSA. LSA adapts frozen encoder features through a local residual transformation on the original patch grid. The spatial operation mixes neighboring patches before channel projection, while the residual path retains the original content. This supplies a locally adapted feature at each position without replacing the grid by a pooled study representation. For encoder grid X and validity mask m, let $\mathcal{M}_m$ clear invalid positions. Let DW and PW denote depthwise and pointwise convolutions. With channel-wise affine layer normalization, LSA is

$$\begin{aligned} D_L &= \mathrm{GELU}(\mathrm{DW}_{3\times3}(\mathcal{M}_m(\mathrm{LN}_{768}(\mathcal{M}_m(X))))), \\ L &= \mathcal{M}_m(\mathrm{LN}_{512}(\mathrm{PW}_{768\to512}(\mathcal{M}_m(X+\gamma D_L)))) . \end{aligned} \tag{13}$$

The second input mask removes affine-normalization offsets at invalid positions before spatial convolution. Both convolutions are bias-free; the depthwise convolution uses padding of one. The shared trainable scalar γ starts at 0.1. Spatial convolution uses the original grid, before valid tokens are compacted. All layer normalization operators use population variance and $\epsilon = 10^{-5}$; N_0 has no affine parameters and is not ℓ_2 normalization.

CRM. CRM augments each frontal-image feature with a reference from the opposite side of the same image. A learned axis gives the nominal reflection, and a bounded local offset adjusts the sampling position. The local-minus-reference feature passes through a spatial transformation whose contribution is controlled by a feature-dependent gate. This modulation supplies within-image comparative information before cross-examination matching; it does not assign a temporal state or assume that the opposite side is healthy. Reference-validity masking preserves the original local feature where sampling is invalid, and nonfrontal views bypass the operation.

FSR. FSR transforms the adapted visual features through a frozen pretrained block, with trainable projections that connect the visual feature width to the block interface. Bidirectional attention lets valid patches interact within an image while preserving one output per input position. Freezing the block constrains which weights are optimized but still allows the surrounding adapters to be trained by backpropagation. Set G to the CRM output for PA/AP images and to L for other views. FSR retains each valid position:

$$F = \mathrm{LN}_{\mathrm{out}}\big(A_{\mathrm{down}}\big(B_{14}^{\mathrm{frozen}}(A_{\mathrm{up}}(\mathrm{LN}_{\mathrm{in}}(G)))\big)\big) . \tag{14}$$

The adapters have widths $512 \to 2048$ and $2048 \to 512$; B_{14} is the fifteenth block, with zero-based index 14, of Llama-3.2-1B (Meta, 2024). Only this block is used. Its weights remain frozen while gradients are propagated to both adapters. Attention is bidirectional within an image; language causal masking and language rotary position embeddings are disabled. There is no cross-image attention or fixed-slot compression. This follows the frozen-block adaptation principle of Tang et al. (2025); using its position-level output for shared longitudinal matching is the present interface choice. The FSR deletion ablation also removes its dedicated adapters and therefore does not isolate the pretrained block alone.

The study means g_C, g_P average all valid patches, rather than equally weighting per-image means. The current visual source contains all current patches and their mean, regardless of prior-view compatibility. Cross-study attention uses separate bias-free $512 \to 512$ query/key projections, joint normalization over all compatible prior patches, and original prior features as values, without extra value/output projections. The relation MLP has widths $1024 \to 512 \to 512$ with GELU between its linear layers.

A.2.1 CRM GEOMETRY AND IMPLEMENTATION ALGORITHM

The following definitions specify the reflection geometry, local offsets, feature descriptors, and validity masking used in CRM.

Axis and reflection. Patch centers use $p_i = (x_i, y_i) \in \{0, \ldots, 36\}^2$, with $\bar{x}_i = (x_i - 18)/18$ and $\bar{y}_i = (y_i - 18)/18$. Two zero-initialized scalars, shared across frontal images and studies, define

$$\Delta = 4\tanh\theta_\Delta, \qquad \kappa = 2\tanh\theta_\kappa. \tag{15}$$

The axis is $x = 18 + \Delta + \kappa\bar{y}$, equivalently $x = ay + b$ with $a = \kappa/18$ and $b = 18 + \Delta - \kappa$. We use Euclidean reflection in this line, not independent horizontal reflection within each row:

$$q_i = \frac{x_i - ay_i - b}{1 + a^2}, \qquad p_i^{\mathrm{mir}} = (x_i - 2q_i,\ y_i + 2aq_i). \tag{16}$$

Table 4: CRM algorithm. Geometry and validity criteria are defined in Eqs. 15–19.

Input: LSA grid $L \in \mathbb{R}^{37\times37\times512}$, patch mask m, and view.
Clear invalid entries of L; return L immediately for a nonfrontal view.
Construct the patch-center grid p and normalized coordinates $\bar{p}$.
Obtain shared bounded Δ, κ; form line coefficients a, b.
Reflect p across $x = ay + b$ to obtain p^{mir}.
Form six-dimensional γ^6 from $p, p^{\mathrm{mir}}, \Delta, \kappa$.
Predict bounded local offsets $\delta = \frac{1}{2}\tanh(\mathrm{MLP}_{\mathrm{off}}([L; \gamma^6]))$.
Set $p^{\mathrm{ref}} = p^{\mathrm{mir}} + \delta$ and determine reference-validity flags u.
Replace invalid sampling coordinates by a safe point and bilinearly sample L.
Clear invalid references in C^{opp}; form $h = m \wedge u$.
Form γ^4 from normalized reference coordinates and scaled offsets.
Compute $A = \sigma(\mathrm{MLP}_{\mathrm{gate}}([L; C^{\mathrm{opp}}; \gamma^4]))$.
Clear invalid differences: $D = \mathcal{M}_h(L - C^{\mathrm{opp}})$.
Apply DW, LN, GELU, and PW in sequence to obtain U.
Return $G = \mathcal{M}_m(L + \mathcal{M}_h(A \odot U))$.

Thus the tilt bound concerns displacement of the axis over normalized image height, rather than a slope of two in patch coordinates.

Local offset and descriptors. Write $\bar{p} = (p - (18, 18))/18$ for normalized coordinates. The six-dimensional pre-offset descriptor is $\gamma_i^6 = (\bar{x}_i, \bar{y}_i, \bar{x}_i^{\mathrm{mir}}, \bar{y}_i^{\mathrm{mir}}, \Delta/4, \kappa/2)$. It depends only on quantities available before offset prediction. With a biased $518 \to 128 \to 2$ GELU MLP,

$$\delta_i = \tfrac{1}{2}\tanh(\mathrm{MLP}_{\mathrm{off}}([L_i; \gamma_i^6])), \qquad p_i^{\mathrm{ref}} = p_i^{\mathrm{mir}} + \delta_i. \tag{17}$$

Each offset component is bounded by half a patch. The four-dimensional gate descriptor is $\gamma_i^4 = (\bar{x}_i^{\mathrm{ref}}, \bar{y}_i^{\mathrm{ref}}, 2\delta_{i,x}, 2\delta_{i,y})$. These are scalar geometry features, not additional image tokens. Descriptor components are not clipped, so an out-of-image reflection remains distinguishable before validity masking.

Sampling and masking. Bilinear sampling reads the same image's masked LSA grid at p_i^{ref}, using normalized sampler coordinates $(2x_i^{\mathrm{ref}}/36 - 1, 2y_i^{\mathrm{ref}}/36 - 1)$ and `align_corners=True`. A reference is valid only if its coordinates are finite, lie in $[0, 36]^2$, and every in-bounds grid point with nonzero interpolation weight has $m = 1$. Zero-weight neighbors impose no validity requirement. Invalid sampling coordinates receive a safe placeholder and their reference values are cleared. The same safe coordinates are used in the gate descriptor at invalid positions. For the reference-validity flag u_i, define $h_i = m_i \wedge u_i$ and denote the sampled feature by C_i^{opp}. The biased gate MLP has widths $1028 \to 256 \to 512$, with hidden GELU:

$$A_i = \sigma(\mathrm{MLP}_{\mathrm{gate}}([L_i; C_i^{\mathrm{opp}}; \gamma_i^4])), \qquad D = \mathcal{M}_h(L - C^{\mathrm{opp}}). \tag{18}$$

The difference path and residual output are

$$U = \mathrm{PW}_{1\times1}(\mathrm{GELU}(\mathrm{LN}_{512}(\mathrm{DW}_{3\times3}(D)))), \qquad G = \mathcal{M}_m(L + \mathcal{M}_h(A \odot U)). \tag{19}$$

Here depthwise convolution uses padding of one and no bias; the pointwise convolution includes a bias. Masking occurs before spatial mixing and again before residual addition. Consequently an invalid reference preserves a valid local feature, even if neighboring differences or convolution biases are nonzero. Nonfrontal images return $\mathcal{M}_m(L)$ directly.

A.3 Finding keys, eligibility, and text initialization

A canonical key comprises a finding concept, anatomical scope, and laterality when available. The same ordered dictionary indexes state-head rows, parsed assertions, supervision, temporal records, and classifier-derived identities. A missing anatomical field is not silently replaced by a side-specific key. Parser uncertainty/conflicts, incomplete scope, and missing token spans remain separate validity fields. The exact dictionary and parser version are dataset-preprocessing artifacts; architecture formulas do not determine their contents.

Eligibility follows three levels. First, select the latest earlier same-patient study in the same partition having usable images and a report, without inspecting the current target report. Second, test compatible views in that selected study. Third, apply finding-specific assertion, scope, and token rules. If the latest selected prior has only AP and the current study only PA, the temporal source is empty even when an older PA exists. Missing spans and views do not cause fallback to another prior. A valid OLD key needs a prior-positive assertion, usable positive-span tokens, and a nonempty relation set. A NEW candidate needs a valid prior state of negative or not mentioned, complete scope, and an eligible comparison. Prior uncertainty or unresolved assertion conflict admits neither path. Current-state filtering affects S and NEW emission, not OLD eligibility.

The state order is positive, negative, uncertain, and not mentioned. The State Head is one $512 \to 4K$ linear map for K keys; its outputs are reshaped key-first. State type embeddings have three shared rows, for positive, negative, and uncertain states. The key's identity is N_0 of the difference between the positive and negative classifier weights, with no classifier bias, additional identity table, or gradient detachment. Argmax ties use the first class in the stated order. NEW uses one $1024 \to K$ linear map and sigmoid; the emission threshold is $\tau = 0.8$, shared by prior NEG and NM states. A key-to-row map retrieves the positive row for a key from compacted S. NEW adds e_{new} once, without another identity addition or outer normalization, and does not modify S in place.

RadGraph parsing and BiomedBERT encoding both receive the original prior report; the parser does not encode text for BiomedBERT. Character-offset overlap maps each positive assertion span to original subword positions, excluding special tokens. Repeated or overlapping mentions contribute each original position once. If an encoding window repeats a position, its hidden vectors are averaged before projection; a configured truncation must mark excluded spans unusable rather than inventing token features. The checkpoint, tokenizer, and window settings are recorded with preprocessing. For the unique retained position set $\mathcal{U}_c$, text projection $W_B \in \mathbb{R}^{768\times512}$ and bias b_B give

$$u_t = N_0(H_t W_B + b_B), \qquad h_c^0 = N_0\left(\frac{1}{|\mathcal{U}_c|}\sum_{t\in\mathcal{U}_c} u_t\right). \tag{20}$$

A key with an empty retained-position set is ineligible for OLD; it is not reassigned to NEW. This initialization contains no visual pre-read and no current-report token.

A.4 Post-refinement prediction and record assembly

The final branch order is fixed. The final branch states are concatenated and projected by a biased $2048 \to 512$ map to r_c. A trusted elapsed-day interval may affect this post-refinement input only:

$$t_{PC} = m_\Delta \tanh(\log(1+\Delta_{\text{days}})W_\Delta + b_\Delta), \qquad r_c^{\text{time}} = r_c \odot (1 + t_{PC}). \tag{21}$$

The interval map has width $1 \to 512$ and zero-initialized weights/bias. The flag m_Δ is zero when interval metadata is unavailable or the optional adjustment is disabled. Invalid intervals are rejected before the logarithm; study-order differences are not converted to days. The activation setting, interval source, and validity coverage are run-level configuration fields. This adjustment does not modify R, expert states, current-state prediction, NEW scores, or the usage head directly.

A two-class softmax of r_c^{time} gives p_c^{root} for persistence/resolution; a three-class softmax gives p_c^{sub} for stable/improved/worsened. The four transition weights are

$$\pi_{c,j} = p_{c,\text{persistent}}^{\text{root}} p_{c,j}^{\text{sub}} \ (j=0,1,2), \qquad \pi_{c,3} = p_{c,\text{resolved}}^{\text{root}}. \tag{22}$$

They sum to one, follow all branch computations, and are not masked by current argmax states. No separate branch-role loss is used. The usage head is shared across branches and learns contribution, not a calibrated probability that the evidence is correct. OLD log-usage is computed in a numerically stable form as $\text{LSE}_j(\log \pi_{c,j} + \log \sigma(N_0(z_{c,j})w_g + b_g))$; NEW uses log-sigmoid of its raw score. Temporal values are never multiplied by usage before decoder attention. OLD and NEW are concatenated along the record axis; content, key identifiers, branch identifiers, validity, and log-usage receive the same ordering permutation.

A.5 Decoder visibility and source normalization

Each decoder layer has width 512, eight heads, and feed-forward hidden width 2048. Four learned registers act as report-internal states, not cross-examination memory. Registers may read registers

and external sources, but never report tokens. A text position reads the registers and its causal text prefix. Registers are updated in every layer except the last, where their states serve only as keys and values. These rules prevent access to future target words through a register. Text self-attention uses rotary position embeddings (Su et al., 2024).

Within a layer, the three sources share one query projection but have separate key, value, and output projections. Query inputs and source key/value inputs have their own normalizations. Log-usage is added to temporal logits once, before both within-source softmax and source-score computation. The current sources use their valid entry counts for source normalization; the temporal source uses a denominator of one. Three learned source biases are shared across heads. An empty source has zero context and is excluded from source competition, without evaluating an empty softmax.

For each nonempty source, let $Z_{a,b} = \sum_i e^{\ell_{a,i,b}}$ and $D_b = \sum_{a \in \mathcal{A}} e^{\delta_a} Z_{a,b}/n_a$. Multiplying the source softmax and within-source softmax yields the effective value weight

$$\omega_{a,i,b} = \frac{e^{\delta_a} e^{\ell_{a,i,b}}}{n_a D_b}. \tag{23}$$

For temporal entries, $e^{\ell_{T,i,b}} = e^{q_b k_{T,i,b}^\top / \sqrt{d_h}} g_i$ contains one usage factor. The shared normalizer does not introduce another value-side multiplication by g_i. Uniformly reducing every temporal usage therefore reduces temporal source strength, including when only one temporal record is present. Keeping $n_T = 1$ also lets additional temporal records increase that strength; invariance to record count is not claimed.

Head-wise source weights multiply their corresponding contexts before source-specific output projection. Each sublayer follows $x \leftarrow x + \mathrm{Dropout}(\mathrm{Sublayer}(\mathrm{LN}(x)))$: normalization precedes the sublayer, and the residual addition follows it. This order applies to causal self-attention, source-aware cross-attention, and the feed-forward sublayer. The final vocabulary projection shares weights with token embeddings. A single temporal record can still change its source contribution even though its within-source weight is one. Retaining the temporal denominator of one means that source strength is not invariant to temporal-record count.

A.6 Optimization and model selection

Table 5: Training configuration for ReCER.

Parameter	ReCER
Optimizer	AdamW
Peak learning rate	1×10^{-4}
Batch size	16
Epochs	30
Learning-rate scheduler	Linear warmup + cosine decay
Warmup ratio	0.05
Precision	BF16
Gradient clipping	1.0
Weight decay (matrix parameters)	0.01

The training configuration is shared across the three datasets. RAD-DINO and the FSR core are frozen. For longitudinal datasets, BiomedBERT is frozen and the active adapters, classifiers, experts, transition/usage heads, and decoder are optimized jointly. In the current-image-only IU configuration, BiomedBERT is not instantiated, temporal supervision is disabled, and inactive historical modules are not optimized. Matrix parameters use weight decay of 0.01. Biases, normalization parameters, and geometric scalars are excluded from weight decay. Gradient clipping applies to the accumulated gradient norm. BF16 autocasting is combined with 32-bit floating-point parameters, optimizer states, and reductions.

An optimizer update uses up to 16 real current-study targets, with each study and its selected prior forming a microbatch unit. The specified seed is 17. Checkpoint selection uses validation CE-F1,

with an earlier checkpoint preferred on a tie. Greedy decoding stops at the end-of-sequence token or the configured length limit.

A.7 Supervision domains

Current and, when available, prior reports provide four-state labels under the same canonical-key mapping. They supervise unfiltered logits, including not-mentioned states. Prior-image state prediction is auxiliary training supervision and never replaces assertions parsed from the selected prior report at inference.

Positive, explicitly negated, and uncertain assertions receive the positive, negative, and uncertain state labels, respectively; a mapped key without a mention receives the not-mentioned label. Contradictory assertions or ambiguous anatomy/laterality are marked invalid for the affected scope. Unknown annotation validity is not treated as an additional clinical state. Scope completeness concerns the parser's ability to bind a key, not proof that an unmentioned abnormality is absent.

Temporal labeling additionally needs reference resolution. A comparative statement is mapped to its finding key, its direction, and the study it explicitly or unambiguously references. Its label is usable only when that reference is the selected prior; ambiguous references, comparisons to a different study, and contradictory directions are masked. Explicitly reference-bound "new" assertions can supply emergence labels. RadGraph assertion labels alone do not supply this temporal adjudication, and silent report-to-report changes are insufficient.

Transition labels must concern the same finding key and selected prior. An unavailable or unreliable transition label suppresses its loss, not an otherwise valid OLD record. NEW positives require explicit, reliably reference-bound evidence of emergence. A negative-to-positive or not-mentioned-to-positive report change alone does not provide that evidence. Reliable persistent findings can supply NEW negatives even though they cannot enter the NEW output branch. Resolved findings are not automatically assigned negative NEW labels. Thus, the NEW supervision domain is independent of candidate filtering, current argmax decisions, and emission thresholds.

Report loss is averaged over effective target-token occurrences. The auxiliary losses are averaged over unique study–key units for assertion prediction and unique current-study–prior-study–key units for transition and NEW prediction. Repeated occurrences are averaged within each optimizer update before unique units are averaged. Every task uses its own valid denominator, and an empty task contributes zero. Current and prior assertion rows form one deduplicated task. All four loss coefficients are one; no extra branch-role, root/subclass, or usage-calibration objective is added.

A.8 Evaluation metrics

BLEU-4 measures modified n-gram precision up to order four with a brevity penalty. METEOR combines aligned unigram precision and recall with a fragmentation penalty. ROUGE-L measures sequence similarity using the longest common subsequence. These language metrics compare generated reports with reference reports.

Clinical efficacy (CE) measures agreement between clinical labels extracted from generated and reference reports. Precision, recall, and F1 use true positives (TP), false positives (FP), and false negatives (FN): $P = TP/(TP+FP)$, $R = TP/(TP+FN)$, and $F1 = 2TP/(2TP+FP+FN)$. ReCER uses micro-averaging, pooling these counts across evaluated report–label pairs. Higher values indicate better performance for all reported metrics.

B Supplementary Configuration Analyses

B.1 Complete MIMIC-CXR Parameter Sensitivity Results

Table 6: Parameter sensitivity results on MIMIC-CXR.

Factor	Setting	BLEU-4	METEOR	ROUGE-L	CE-P	CE-R	CE-F1
Experts	$E = 3$	0.168	0.178	0.318	0.572	0.472	0.517
	$E = 4$	0.178	0.190	0.329	0.576	0.479	0.523
	$E = 5$	0.177	0.190	0.328	0.574	0.477	0.521
	$E = 6$	0.169	0.174	0.315	0.566	0.475	0.517
Updates	$r = 1$	0.172	0.175	0.322	0.574	0.474	0.519
	$r = 2$	0.178	0.190	0.329	0.576	0.479	0.523
	$r = 3$	0.169	0.180	0.314	0.568	0.473	0.516
Decoder	$D = 3$	0.167	0.178	0.314	0.574	0.468	0.516
	$D = 4$	0.178	0.190	0.329	0.576	0.479	0.523
	$D = 5$	0.173	0.174	0.322	0.573	0.472	0.518

Table 6 reports the complete MIMIC-CXR sensitivity results for expert count, update count, and decoder depth, with Figure 3 visualizing the same values. Across the tested settings, the default configuration of four experts, two update rounds, and four decoder layers achieves the highest or tied-highest performance on all reported metrics. Non-default configurations exhibit different language-metric trade-offs: three updates yield higher METEOR than one update (0.180 vs. 0.175) but lower BLEU-4 and ROUGE-L, whereas five decoder layers yield higher BLEU-4 and ROUGE-L than three layers but lower METEOR. Overall, the selected default configuration provides the strongest performance across the reported clinical and language metrics.

B.2 CheXpert Plus module ablations

Table 7: Ablation study of ReCER on CheXpert Plus. Column bests are bold, including ties.

Configuration	BLEU-4	METEOR	ROUGE-L	CE-P	CE-R	CE-F1
Full	**0.134**	**0.158**	**0.323**	**0.388**	**0.327**	**0.355**
w/o experts	0.122	0.149	0.309	0.361	0.325	0.342
w/o CRM	0.126	0.153	0.315	0.380	0.301	0.336
w/o FSR	0.128	0.154	0.318	0.373	0.319	0.344
w/o LSA	0.131	0.156	0.320	0.383	0.321	0.349
w/o CRM+FSR+LSA	0.124	0.151	0.312	0.364	0.300	0.329
w/o NEW head	0.127	0.154	0.317	0.365	0.320	0.341
w/o registers	0.132	0.157	0.321	0.387	0.322	0.352

Expert replacement yields CE-F1 of 0.342, compared with 0.355 for Full. Precision changes from 0.388 to 0.361 and recall from 0.327 to 0.325. Current sources and one ordinary relation read remain, so this is not a current-only configuration. Joint visual replacement yields CE-F1 of 0.329; CRM, FSR, and LSA replacements yield 0.336, 0.344, and 0.349. NEW-head replacement yields 0.341 and register removal 0.352. Full leads all metrics, gaining 0.013/0.026 CE-F1 over expert/joint visual replacement and supporting the complete design across both longitudinal datasets.

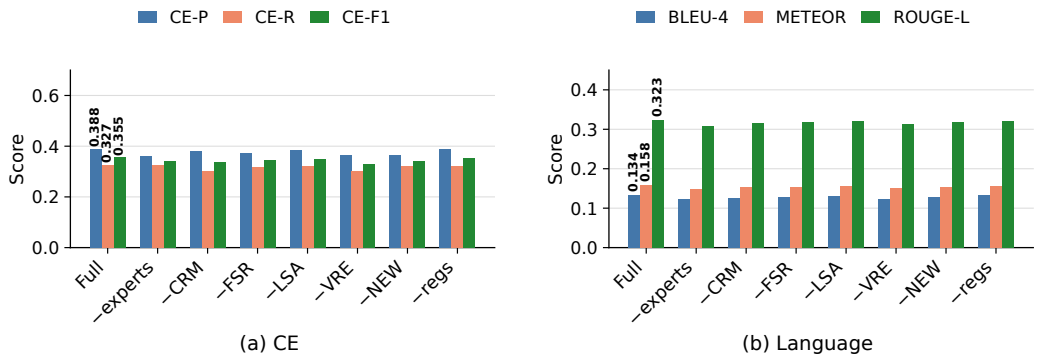


Figure 5: Ablation results of ReCER on CheXpert Plus. The left and right panels show clinical and language metrics, respectively, for the results in Table 7. Best values per metric are annotated in bold (ties included).

B.3 CheXpert Plus parameter sensitivity

Table 8: Parameter sensitivity on CheXpert Plus.

Configuration	BLEU-4	METEOR	ROUGE-L	CE-P	CE-R	CE-F1
Expert count						
$E = 3$	0.123	0.146	0.307	0.380	0.319	0.347
$E = 4$	0.134	0.158	0.323	0.388	0.327	0.355
$E = 5$	0.130	0.153	0.319	0.384	0.324	0.352
$E = 6$	0.125	0.142	0.310	0.372	0.324	0.346
Expert update rounds						
$r = 1$	0.128	0.149	0.308	0.383	0.322	0.350
$r = 2$	0.134	0.158	0.323	0.388	0.327	0.355
$r = 3$	0.124	0.143	0.313	0.384	0.315	0.346
Decoder depth						
$D = 3$	0.126	0.145	0.307	0.380	0.307	0.340
$D = 4$	0.134	0.158	0.323	0.388	0.327	0.355
$D = 5$	0.122	0.153	0.311	0.376	0.317	0.344

Configurations with four, five, three, and six experts achieve CE-F1 scores of 0.355, 0.352, 0.347, and 0.346, respectively. Three experts have higher precision but lower recall than six. Three-expert BLEU-4/METEOR/ROUGE-L scores are 0.123/0.146/0.307, versus 0.125/0.142/0.310 for six experts; the five-expert scores are 0.130/0.153/0.319.

Configurations with one, two, and three updates achieve CE-F1 scores of 0.350, 0.355, and 0.346, respectively. One-update language scores are 0.128/0.149/0.308, versus 0.124/0.143/0.313 for three. Configurations with three, four, and five decoder layers achieve CE-F1 scores of 0.340, 0.355, and 0.344, respectively. Five layers have higher METEOR/ROUGE-L than three (0.153/0.311 versus 0.145/0.307), but lower BLEU-4 (0.122 versus 0.126). Interfaces are defined in Appendix D. Four experts, two updates, and four decoder layers lead all CheXpert Plus metrics, supporting the same architecture across both longitudinal datasets.

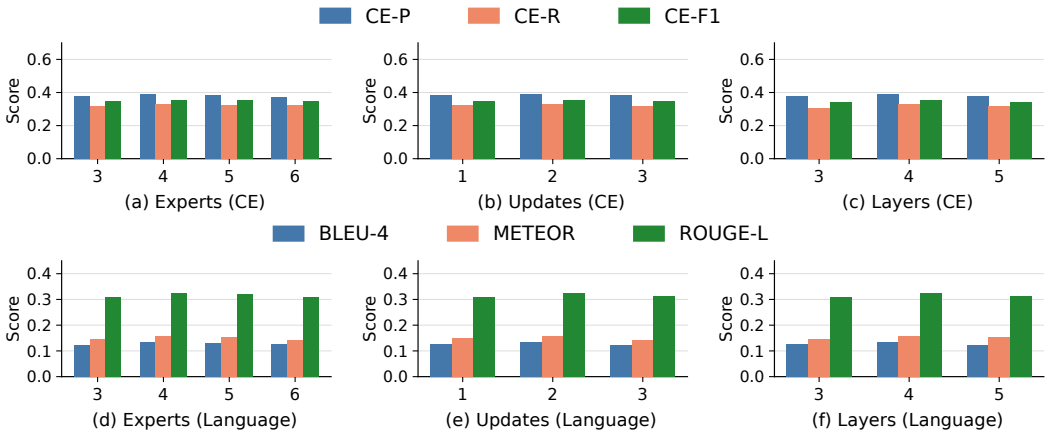


Figure 6: Parameter sensitivity on CheXpert Plus. Columns vary expert count, update count, and decoder depth. The upper and lower rows show clinical and language metrics, respectively. Axes show absolute scores; full values appear in Table 8.

B.4 IU X-Ray ablations without history

Table 9: Ablation study of ReCER on IU X-Ray. Column bests are bold, including ties.

Configuration	BLEU-4	METEOR	ROUGE-L	CE-P	CE-R	CE-F1
Full	**0.211**	**0.234**	**0.397**	0.557	**0.598**	**0.576**
w/o experts	—	—	—	—	—	—
w/o CRM	0.205	0.231	0.390	0.547	0.559	0.553
w/o FSR	0.206	0.232	0.391	0.538	0.588	0.562
w/o LSA	0.208	0.233	0.394	**0.559**	0.575	0.567
w/o CRM+FSR+LSA	0.202	0.229	0.387	0.536	0.560	0.548
w/o NEW head	—	—	—	—	—	—
w/o registers	0.210	**0.234**	0.395	0.555	0.587	0.571

Without a valid prior, only current finding and visual sources are active. OLD experts and NEW are therefore inapplicable as direct runtime ablations; dashes denote inapplicability, not a measured zero effect. Joint visual replacement yields CE-F1 of 0.548. CRM, FSR, and LSA replacements yield 0.553, 0.562, and 0.567. LSA removal has slightly higher precision than Full (0.559 versus 0.557) but lower recall (0.575 versus 0.598). Register removal yields CE-F1 of 0.571 with unchanged METEOR of 0.234. These records concern current-image representation and expression. Full achieves the highest CE-F1 (0.576), exceeding joint visual replacement by 0.028 and supporting the visual modules' contribution without prior examinations.

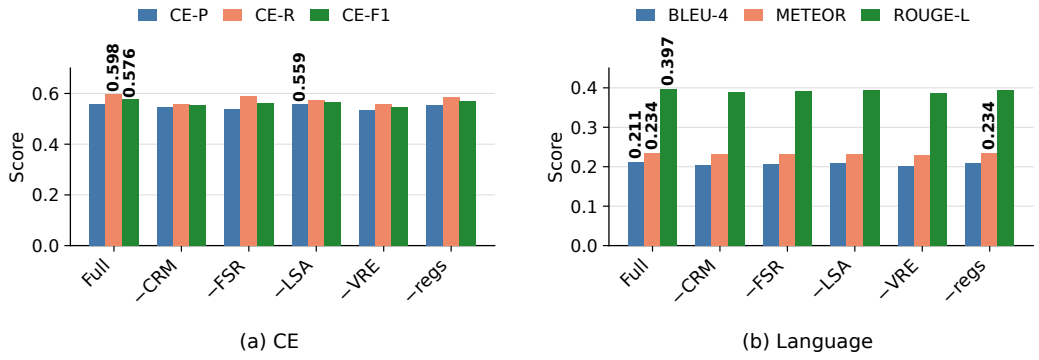


Figure 7: Ablation results of ReCER on IU X-Ray. The left and right panels show clinical and language metrics, respectively, for the results in Table 9. Best values per metric are annotated in bold (ties included).

B.5 IU X-RAY DECODER DEPTH

Table 10: Decoder-depth comparison on IU X-Ray.

Configuration	BLEU-4	METEOR	ROUGE-L	CE-P	CE-R	CE-F1
Decoder depth						
$D = 3$	0.202	0.220	0.390	0.552	0.578	0.565
$D = 4$	0.211	0.234	0.397	0.557	0.598	0.576
$D = 5$	0.200	0.229	0.382	0.551	0.591	0.570

Configurations with three, four, and five layers achieve CE-F1 scores of 0.565, 0.576, and 0.570, respectively. Three-layer BLEU-4/METEOR/ROUGE-L scores are 0.202/0.220/0.390, versus 0.200/0.229/0.382 for five layers. Five layers therefore have higher METEOR and CE-F1 than three, but lower BLEU-4/ROUGE-L. This comparison concerns current-image report construction, not temporal refinement, because the historical pathway is inactive. Four decoder layers lead all tested metrics (CE-F1: 0.576), supporting the selected depth for both current-image-only and longitudinal reporting.

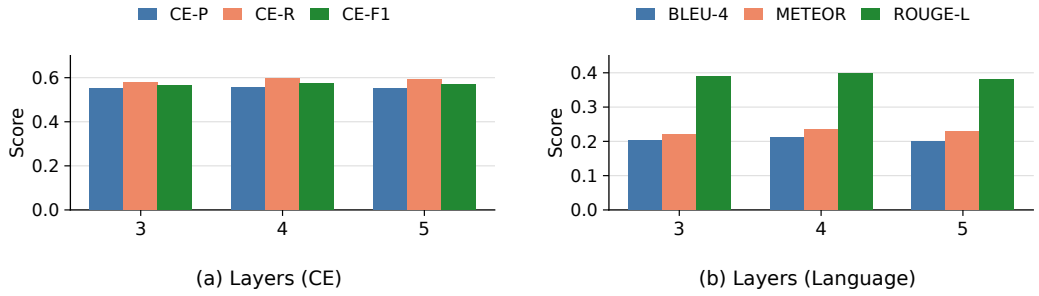


Figure 8: Parameter sensitivity on IU X-Ray. The panels show clinical and language metrics across decoder depths. Axes show absolute scores; full values appear in Table 10.

C Parameter Count

Table 11: Module-wise parameter counts for ReCER (in millions). The Trainable columns report parameters optimized for report generation; the Total columns also include frozen parameters. MIMIC-CXR and CheXpert Plus share the same configuration.

	MIMIC-CXR / CheXpert Plus		IU X-Ray	
Module	Trainable	Total	Trainable	Total
State type	0.00	0.00	0.00	0.00
NEW type	0.00	0.00	–	0.00
State Head	0.32	0.32	0.08	0.08
NEW Head	0.16	0.16	–	0.04
Relation query/key	0.52	0.52	–	0.52
Relation MLP and norm	0.79	0.79	–	0.79
Text projection	0.39	0.39	–	0.39
OLD shared recurrent reader	0.33	0.33	–	0.33
OLD expert stable	2.10	2.10	–	2.10
OLD expert improved	2.10	2.10	–	2.10
OLD expert worsened	2.10	2.10	–	2.10
OLD expert resolved	2.10	2.10	–	2.10
OLD fusion, transition, usage	1.05	1.05	–	1.05
Decoder	38.78	38.78	38.78	38.78
RAD-DINO	0.00	86.58	–	86.58
LSA	0.40	0.40	0.40	0.40
CRM	0.73	0.73	0.73	0.73
FSR frozen block	0.00	60.82	–	60.82
FSR trainable adapters	2.10	2.10	2.10	2.10
BiomedBERT	0.00	108.89	–	–
Total	53.99	310.28	42.10	201.03

Shared tensors are counted once; values below 0.005M round to 0.00M. The labels stable, improved, worsened, and resolved identify the ordered transition-weight slots; they do not denote independently supervised or verified clinical expert roles.

Table 11 details the parameter inventory. In the IU columns, a dash in Trainable denotes no optimized parameters, while a dash in Total denotes an absent module; rounded nonzero values remain 0.00. MIMIC-CXR and the matching CheXpert Plus configuration use 310.28M total parameters, including 53.99M trainable and 256.29M frozen parameters. The CheXpert counts follow the shared architecture rather than a separate profiling run. IU uses 201.03M total, 42.10M trainable, and 158.93M frozen parameters; it omits BiomedBERT and freezes unused historical modules. The 38.78M decoder is included in each total, with tied input and output embeddings counted once.

Table 12: Comparison of report generation parameter counts across models (in millions). MLRG's count is taken from Table 7 in the extended version of PriorRG, which also reports PriorRG's total (Liu et al., 2026a).

Method	Year	Count scope	Report generation
RECAP	2023	Trainable	160.05
MLRG	2025	Total	296.00
PriorRG	2026	Total	335.00
ReCER	2026	Total / trainable	310.28 / 53.99

RECAP reports trainable parameters (Hou et al., 2023); ReCER uses the MIMIC-CXR/CheXpert Plus configuration, with total and trainable counts listed in that order. PriorRG's count corresponds to Stage 2.

Table 12 compares report-generation parameter counts while retaining the source-specific counting scope. ReCER's 310.28M total describes its report-generation inventory, whereas 53.99M parameters are optimized for that task. Only 17.4% are trainable; the total is 24.72M (7.4%) below PriorRG under the stated counting scopes.

D Ablation Definitions and Parameter Interfaces

D.1 Relation-expert replacement

Each ablation variant is trained independently with the same split, model-selection rule, and decoding protocol; pruning a trained full model would be a different intervention. Expert replacement retains prior-text initialization and local relations but performs one ordinary read with the shared projections:

$$b_c = \sum_{i \in \mathcal{I}_R} \mathrm{softmax}_{i \in \mathcal{I}_R} \left(\frac{[N_0(h_c^0) W_Q^E](K_i^E)^\top}{\sqrt{d_q}} \right) R_i. \tag{24}$$

The same readout fills all four posterior slots. Branch identities, state projections, expert MLPs, and recurrent updates are removed. Transition and usage heads, finding identity, and current/NEW paths remain. Identical slots make content and usage independent of the transition mixture, although transition supervision can still train its head. This tests the coupled replacement, not an isolated capacity-matched MLP or complete history removal.

D.2 Visual modules, NEW head, and registers

Joint visual replacement. RAD-DINO, the 768-to-512 projection, output normalization, and valid-position metadata remain. Local adaptation, CRM, and FSR are bypassed for both studies. Replacement features serve current states, NEW summaries, cross-study matching, and the visual source.

Individual visual replacements. Removing CRM bypasses contralateral sampling and gated differences for frontal images, while retaining FSR and cross-study attention. Removing FSR bypasses its frozen core and adapters, preserving width and position metadata. Removing LSA disables its local residual but retains channel projection and normalization. Joint effects cannot be obtained by summing individual score differences.

NEW head. The learned head, fixed NEW emission threshold, and NEW loss are removed. Candidates passing the prior-based and current-positive rules are emitted with unit usage. Content is still formed by retrieving a positive current record and adding the NEW type embedding. Pair/scope validity and OLD/NEW exclusivity remain. Learned selection, attenuation, and supervision change together.

Registers. Register states, updates, masks, and caches are removed. Causal text attention, source-aware cross-attention, feed-forward layers, and text positions remain.

D.3 Four-slot adaptation for expert count

The number of expert branches varies, while the posterior retains four ordered slots and transition classes. Branch MLP width and shared read/state projections remain unchanged. Final branch states are mapped to four slots by interval overlap:

$$\widetilde{z}_k = \sum_{j=0}^{E-1} 4 \left| \left[\frac{k}{4}, \frac{k+1}{4} \right] \cap \left[\frac{j}{E}, \frac{j+1}{E} \right] \right| z_j. \tag{25}$$

Here E is expert count, k indexes slots, j indexes branches, and vertical bars measure interval length. Each slot's weights sum to one; every branch contributes. Four experts give the identity mapping. The adaptation adds no learned parameters, but changing count alters capacity and mixing. Slot indices do not establish independently supervised clinical roles.

D.4 Update count, decoder depth, and essential controls

Update-count comparisons fix four experts and reuse branch parameters across steps. Relations and projected keys are computed once; transition and usage follow the final update. Additional steps

increase execution depth without adding independently parameterized experts. Each decoder layer retains width 512, eight heads, hidden width 2048, and four registers; registers update in all but the last layer. Decoder depth changes capacity and register-update count together. Layer-count ratios are not measurements of total parameters or end-to-end latency.

E Additional Qualitative Analysis

E.1 CheXpert Plus example

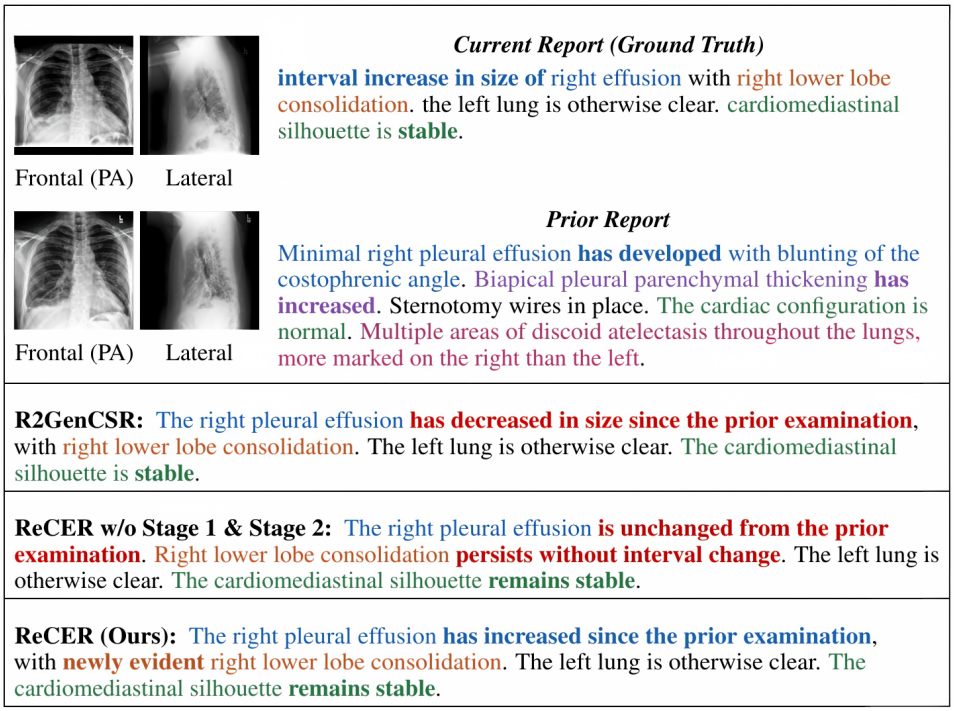
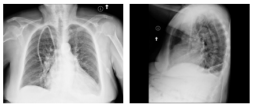


Figure 9: Qualitative comparison of longitudinal report generation on CheXpert Plus. ReCER preserves the effusion increase and adds a temporal qualifier to consolidation beyond the reference wording.

The current reference describes increased right pleural effusion, right lower-lobe consolidation, and a stable cardiomediastinal silhouette. ReCER preserves the reported effusion increase, whereas the baseline-labelled text reverses its direction and the stage-removal variant describes it as unchanged. ReCER further describes the consolidation as "newly evident", making the temporal interpretation more explicit than the reference wording. The displayed reports alone do not establish whether the consolidation is new. A generated report can differ from a reference in descriptive specificity, so additional temporal wording should not automatically be treated as a negative outcome. This example highlights the value of longitudinal reporting that considers both current findings and their relationship to earlier examinations.

E.2 IU X-RAY CURRENT-STUDY EXAMPLE



Frontal Lateral

Current Report (Ground Truth)

Stable cardiomegaly and mediastinal contour. Increased interstitial lung markings are seen, possibly due to volume overload. There is **improved aeration of the lung bases** with small residual left basilar effusion. No XXXX focal consolidation or pneumothorax. Stable tunneled dialysis catheter. Visualized osseous structures appear intact.

R2GenCSR: **Heart size is normal**. Increased interstitial lung markings are seen, possibly due to volume overload. **There is no pleural effusion**. No focal consolidation or pneumothorax is seen. A tunneled dialysis catheter is present. Visualized osseous structures appear intact.

ReCER (Ours; current-only): Cardiomegaly is present. Increased interstitial lung markings may be related to volume overload. A small left basilar pleural effusion is present. No focal consolidation or pneumothorax is seen. A tunneled dialysis catheter is present. Visualized osseous structures appear intact.

Figure 10: Qualitative comparison of report generation on IU X-Ray. Frontal and lateral views come from the same current study; no prior study is used.

The reference describes cardiac enlargement, interstitial markings that may reflect volume overload, a small residual left basilar effusion, and improved basal aeration, while excluding focal consolidation and pneumothorax. The baseline-labelled text under-calls cardiac enlargement and incorrectly negates pleural effusion; the ReCER-labelled text retains more findings and qualifiers. This current-study illustration concerns preservation of report content, not visual verification of temporal change.

F DISCUSSION AND LIMITATIONS

ReCER treats temporal interpretation and historical contribution as related but distinct decisions. A resolved finding may still be relevant to the current report, so current positivity alone is insufficient for deciding whether its history should contribute to generation. Conversely, a newly mentioned finding is not necessarily newly developed. These distinctions motivate separate handling of prior-positive findings and potential emergence, while retaining independent access to current-image evidence. In this formulation, historical information contributes through its interpretation in the current examination, with learned usage modulating the contribution of each temporal record during decoding.

ReCER achieves favorable report-level language quality and clinical label agreement under the reported settings. Differences in evaluation cohorts, label-extraction settings, and CE averaging conventions, including micro and macro averaging, limit direct comparison with literature results. The current metrics do not separately establish the correctness of temporal direction, emergence, or the handling of resolved findings. Future work will harmonize evaluation protocols and incorporate temporal assessments explicitly linked to the selected prior.