# MAKING LONGITUDINAL CONTEXT MATTER FOR RADIOLOGY REPORT GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Longitudinal radiology report generation requires interpreting historical evidence in the context of the current examination. However, static historical representations lack explicit patient-specific adaptation, and simply combining prior images and reports with the current image does not establish their effective contribution. Moreover, full-context cross-entropy supervision can reduce generation loss despite neglecting historical visual evidence or relying primarily on textual patterns in prior reports. We propose PHAM-CECA, an end-to-end framework that couples patient-specific historical representation with contribution-aware optimization. Patient-specific Historical Associative Memory (PHAM) writes prior visual and textual evidence into separate fast-weight memories through patient-specific adaptation. The current image queries both memories to retrieve historical representations tailored to the present examination. The proposed Calibrated Expected Context Advantage (CECA) measures visual, textual, and overall historical contributions by evaluating the same target report under full-history, text-history, image-history, and current-only conditions. PHAM adapts memory parameters per case in the inner loop, while generation supervision and CECA optimize shared parameters in the outer loop. Experiments on the Longitudinal-MIMIC dataset demonstrate the effectiveness of the proposed method.

## 1 INTRODUCTION

Longitudinal radiology report generation uses prior images and reports to contextualize current findings and describe changes over time (Zhu et al., 2023). These sources provide complementary evidence for distinguishing new findings from persistent ones. However, making history available does not ensure its effective use. The generator must retrieve relevant historical evidence and translate it into a measurable contribution to the current report.

Recent methods have advanced longitudinal modeling through several complementary strategies. HERGen (Wang et al., 2024) aggregates examinations through group causal attention. HC-LLM (Liu et al., 2025c) constrains time-shared and time-specific image-report representations, and MLRG (Liu et al., 2025b) combines multi-view and longitudinal evidence through contrastive learning. More recently, TIM (Dong et al., 2026) separates spatial pathology from temporal progression and performs iterative mutual report refinement. These approaches improve the modeling and alignment of temporal evidence. They also raise the question of how to adapt historical representations to each case while explicitly training the generator to benefit from each historical modality.

Effective historical utilization requires both adaptive representation and contribution-aware optimization. Historical modalities provide complementary evidence, but their relevance varies across cases and learning may favor one modality (Peng et al., 2022), motivating separate adaptation and current-conditioned retrieval. Meanwhile, full-context supervision does not explicitly assess historical contributions. Moreover, a larger full-versus-reduced context gap does not necessarily indicate better use of history, because it can also arise when reduced-context predictions become worse. We therefore combine contribution constraints with generation supervision across all contexts.

In this paper, we propose PHAM-CECA, a framework that connects historical representation with its predictive contribution. Inspired by test-time neural associative memory (Behrouz et al., 2026), Patient-specific Historical Associative Memory (PHAM) constructs separate fast-weight memories for prior visual and textual evidence, whose parameters are rapidly adapted to each patient's history. The



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.

current image queries these memories to retrieve historical context relevant to the present examination. Following gradient-based meta-learning (Finn et al., 2017), outer-loop report supervision guides the learning of historical adaptation and retrieval across cases. This design combines modality-specific memory construction with current-conditioned evidence retrieval, aligning historical representations with the report-generation task. The proposed Calibrated Expected Context Advantage (CECA) module optimizes historical utilization by evaluating the same target report under full-history, text-history, image-history, and current-only conditions. Contrasting full-history with ablated-context log-likelihoods isolates the visual, textual, and overall contributions of history. CECA encourages these contributions in expectation while retaining generation supervision across all four contexts to discourage gains obtained by degrading reduced-context predictions. Joint optimization connects PHAM's memory adaptation and retrieval with CECA's contribution-aware objective, enabling the framework to learn both how to represent history and how to use it for report generation.

Our contributions are threefold:

- We propose PHAM, which adapts separate visual and textual memories to each patient and queries them with the current image, producing patient-specific, current-conditioned historical representations.
- We propose CECA, which combines same-target likelihood comparisons across four contexts with generation supervision to optimize expected historical contributions while discouraging reduced-context degradation.
- We demonstrate the effectiveness of our method through theoretical analysis and extensive experiments on Longitudinal-MIMIC (Zhu et al., 2023).

## 2 Related Work

**Longitudinal radiology report generation.** Memory-based report generators include R2Gen (Chen et al., 2020), which maintains a relational memory during decoding, and cross-modal memory networks, which align visual and textual representations through shared memory (Chen et al., 2021). Longitudinal approaches additionally model examinations over time. Zhu et al. (2023) combines the previous image, previous report, and current image for report pre-filling. HERGen (Wang et al., 2024) integrates historical image sequences with a group causal transformer and an auxiliary contrastive objective. MLRG (Liu et al., 2025b) uses multi-view longitudinal contrastive learning and an encoding of absent prior information. LLM-RG4 (Wang et al., 2025) supports different available clinical input contexts, while HC-LLM (Liu et al., 2025c) constrains representations using historical information. Diff-RRG (Yun et al., 2025) uses disease-wise patch differences to condition longitudinal generation. Different to existing methods, our PHAM-CECA uses independent image and report memories, queries both with the current examination, and explicitly optimizes their conditional contribution to the same report target.

**Associative memory and learning through adaptation.** Fast-weight formulations connect attention-like retrieval to transient parameter updates (Schlag et al., 2021), and modern Hopfield networks provide an associative interpretation of attention (Ramsauer et al., 2020). Test-time training layers represent sequence state as the parameters of a model updated by a self-supervised objective (Sun et al., 2024). Titans similarly studies neural memory that learns to retain historical context at test time (Behrouz et al., 2026). In our method, PHAM applies a historical reconstruction update to two small memories for each longitudinal case. Learning their shared initializations by differentiating through the update follows the principle of gradient-based meta-learning (Finn et al., 2017). Our analysis uses the finite-network tangent kernel (Jacot et al., 2018) to connect this update to current-conditioned retrieval and to show how report supervision learns the induced retrieval operator.

**Preference optimization for report generation.** Reward-based methods improve radiology report generation by explicitly optimizing factual quality. Miura et al. (2021) introduces clinical entity and entailment rewards, while Delbrouck et al. (2022) uses RadGraph-based semantic rewards to improve factual correctness. Preference optimization has also been applied to suppress hallucinated references to prior examinations (Banerjee et al., 2024). More recently, RRG-DPO (Liu et al., 2025a) constructs clinically relevant preferences using abnormality conflicts and preserves valid content through sub-preference optimization. These approaches guide generation through feedback on report

content. Contrarily, the proposed CECA focuses on the predictive contribution of historical inputs by comparing the same reference report across four contexts and optimizing expected historical gains alongside generation supervision.

## 3 METHODOLOGY

Given a current image $I_i^{\text{cur}}$ and a prior image-report pair $(I_i^{\text{pre}}, R_i^{\text{pre}})$, the task is to generate a report $R_i^{\text{cur}}$ that describes the patient's current condition and its changes relative to the prior examination. A shared visual encoder independently encodes both images as $H_i^{\text{img}} = E_{\text{vis}}(I_i^{\text{pre}})$ and $H_i^{\text{cur}} = E_{\text{vis}}(I_i^{\text{cur}})$, where $H_i^{\text{img}}, H_i^{\text{cur}} \in \mathbb{R}^{N_Q \times d_V}$. A text encoder extracts features from the prior report, yielding $H_i^{\text{txt}} = E_{\text{txt}}(R_i^{\text{pre}}) \in \mathbb{R}^{N_i^{\text{txt}} \times d_T}$. Here, $N_Q$ and $N_i^{\text{txt}}$ denote the visual and textual token counts, respectively. Figure 1 summarizes the framework.

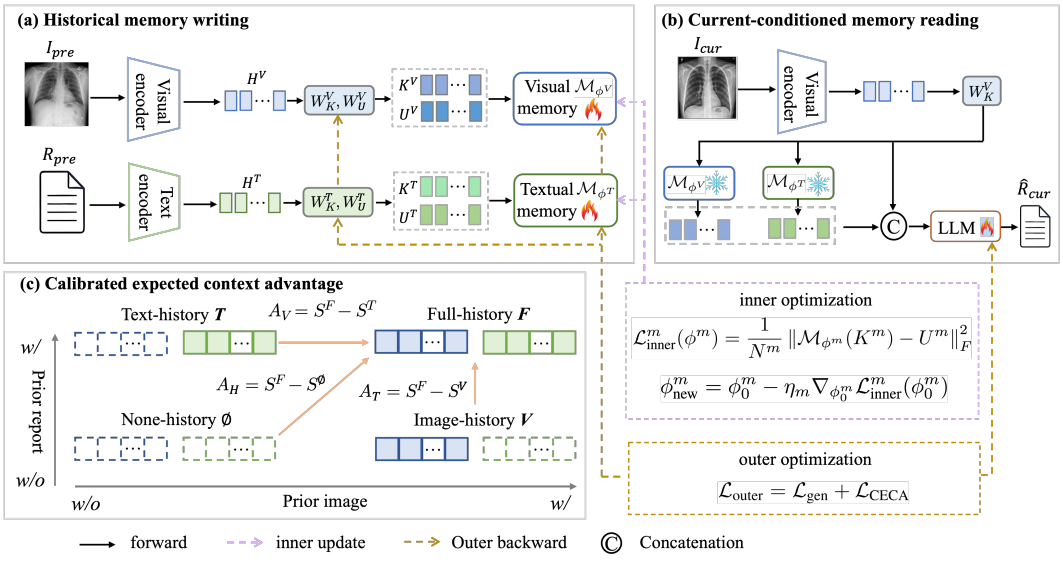



Figure 1: Overview of PHAM-CECA. The framework couples patient-specific historical representation with contribution-aware optimization to promote effective use of visual and textual history in longitudinal report generation.

### 3.1 PATIENT-SPECIFIC HISTORICAL ASSOCIATIVE MEMORY

To capture patient-specific historical evidence, we maintain separate associative memories for prior images and reports and adapt their fast-weight parameters to each case. The current image then queries these memories to retrieve history-aware prompts for report generation. For each modality $m \in \{\text{img}, \text{txt}\}$, we map historical features into key-value pairs of a common memory dimension $d_m$ through modality-specific projections:

$$K_i^m = \text{Norm}(H_i^m W_K^m), \qquad U_i^m = \text{Norm}(H_i^m W_U^m), \tag{1}$$

where Norm denotes rowwise $\ell_2$ normalization and $W_K^m, W_U^m \in \mathbb{R}^{d_L \times d_m}$, $d_L$ denotes the LLM embedding dimension. The resulting keys and values satisfy $K_i^m, U_i^m \in \mathbb{R}^{N_i^m \times d_m}$. Each modality maintains an associative memory $\mathcal{M}_{\phi^m}$, parameterized by a two-layer network $\mathcal{M}_{\phi^m}(x) = W_2^m \,\text{SiLU}(W_1^m x + b_1^m) + b_2^m$ with hidden dimension $d_h$. Starting from a modality-specific initialization $\phi_0^m$ shared across cases, each memory encodes the historical associations of case $i$ through one gradient step on a key-value reconstruction objective:

$$\mathcal{L}_{\text{inner},i}^m(\phi^m) = \frac{1}{N_i^m} \sum_{j=1}^{N_i^m} \left\| \mathcal{M}_{\phi^m}(k_{ij}^m) - u_{ij}^m \right\|_2^2, \quad \phi_i^m = \phi_0^m - \eta_m \nabla_{\phi_0^m} \mathcal{L}_{\text{inner},i}^m(\phi_0^m), \tag{2}$$

where $k_{ij}^m$ and $u_{ij}^m$ denote the $j$-th key and value vectors, respectively, and $\eta_m$ is the modality-specific adaptation rate. This objective guides the inner-loop update of the memory parameters using the historical evidence of case $i$.

To retrieve historical evidence relevant to the current examination, we reuse the visual key projection $W_K^{\text{img}} \in \mathbb{R}^{d_V \times d_L}$ to construct queries from the current visual features:

$$Q_i = \text{Norm}(H_{\text{cur},i}^V W_K^{\text{img}}), \quad Z_i^m = \mathcal{M}_{\phi_i^m}(Q_i), \tag{3}$$

where $H_{\text{cur},i}^V \in \mathbb{R}^{N_Q \times d_V}$. The queries $Q_i \in \mathbb{R}^{N_Q \times d_L}$ read both adapted memories, yielding $Z_i^m \in \mathbb{R}^{N_Q \times d_L}$. Each memory thus produces $N_Q$ historical prompt tokens, conditioned jointly on its historical evidence and the current examination.

### 3.2 Contexts and generation supervision

To support explicit comparisons of historical contributions, we supervise the same reference report under four contexts $F = (I_i^{\text{pre}}, R_i^{\text{pre}}, I_i^{\text{cur}})$, $T = (R_i^{\text{pre}}, I_i^{\text{cur}})$, $V = (I_i^{\text{pre}}, I_i^{\text{cur}})$, and $\emptyset = (I_i^{\text{cur}})$, corresponding to full history, text history, image history, and no history, respectively. Historical image and report features contribute through the retrieved prompts $Z_i^{\text{img}}$ and $Z_i^{\text{txt}}$, respectively, yielding the decoder contexts as:

$$\mathcal{Z}_i^F = [Z_i^{\text{img}}; Z_i^{\text{txt}}; Q_i], \qquad \mathcal{Z}_i^T = [Z_i^{\text{txt}}; Q_i], \qquad \mathcal{Z}_i^V = [Z_i^{\text{img}}; Q_i], \qquad \mathcal{Z}_i^\emptyset = [Q_i]. \tag{4}$$

All contexts use the same instruction and reference report. Teacher-forced supervision is restricted to the current report and its termination token. Let $\bar{R}_i^{\text{cur}} = (R_i^{\text{cur}}, \text{EOS})$ contain $L_i$ supervised tokens, denoted by $r_{it}^{\text{cur}}$. We define the normalized conditional log-likelihood as

$$S_i^c = \frac{1}{L_i} \sum_{t=1}^{L_i} \log p_\theta\left(r_{it}^{\text{cur}} \mid r_{i,<t}^{\text{cur}}, \mathcal{Z}_i^c\right), \qquad c \in \{F, T, V, \emptyset\}. \tag{5}$$

For a batch of $B$ cases, the generation objective is

$$\mathcal{L}_c = -\frac{1}{B} \sum_{i=1}^{B} S_i^c, \qquad \mathcal{L}_{\text{gen}} = \mathcal{L}_F + \alpha_T \mathcal{L}_T + \alpha_V \mathcal{L}_V + \alpha_\emptyset \mathcal{L}_\emptyset, \tag{6}$$

where $\alpha_T, \alpha_V, \alpha_\emptyset > 0$ control the supervision of reduced-history contexts. This objective provides reference-report supervision under every context and remains active during context-advantage optimization.

### 3.3 Calibrated expected context advantage

To quantify the predictive contributions of historical evidence, we define the visual, textual, and overall context advantages as

$$A_{V,i} = S_i^F - S_i^T, \qquad A_{T,i} = S_i^F - S_i^V, \qquad A_{H,i} = S_i^F - S_i^\emptyset. \tag{7}$$

A positive advantage indicates an increase in reference-report log-likelihood when the corresponding historical information is included. We constrain the expected historical contributions, estimated as $\bar{A}_k = \frac{1}{B} \sum_{i=1}^{B} A_{k,i}$, $k \in \{V, T, H\}$. After a warm-up under $\mathcal{L}_{\text{gen}}$, we calibrate the corresponding margins from paired context advantages using one-sided lower confidence bounds (LCB) with nominal coverage $1 - \gamma$:

$$\delta_k = \rho \max\left\{0, \text{LCB}_\gamma(\widehat{A}_k)\right\}, \qquad 0 < \rho \leq 1, \tag{8}$$

where $\widehat{A}_k$ denotes an estimate of the expected context advantage used for margin calibration, $\rho$ controls margin shrinkage. Then, we combine generation supervision with soft constraints on the expected context advantages:

$$\mathcal{L}_{\text{outer}} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{CECA}}, \qquad \mathcal{L}_{\text{CECA}} = \sum_{k \in \{V,T,H\}} \lambda_k [\delta_k - \bar{A}_k]_+, \tag{9}$$

where $[x]_+ = \max(0, x)$ and $\lambda_k \geq 0$. Applying the hinge to mean advantages accommodates variation across cases while encouraging aggregate historical gains. For $\lambda_k > 0$, the penalty is active whenever $\bar{A}_k < \delta_k$. Thus, a zero margin still penalizes negative mean advantages. To remove the direct incentive to decrease a reduced-context score, we couple the generation and advantage weights under the common normalization defined above such that $\alpha_T \geq \lambda_V, \alpha_V \geq \lambda_T, \alpha_\emptyset \geq \lambda_H$. The current report provides outer-loop supervision through $\mathcal{L}_{\text{outer}}$. Its gradients propagate through the decoder, memory reads, and differentiable inner updates to jointly learn the shared memory initializations, projection parameters, and trainable report-generation parameters.

### 3.4 Theoretical Foundations of Historical Context Utilization

We characterize two properties underlying PHAM-CECA: how patient-specific adaptation induces current-conditioned historical retrieval, and how context-advantage optimization promotes historical contributions without directly rewarding degradation of reduced contexts.

**Theorem 1** (Patient-specific adaptation induces task-learned retrieval)**.** *For modality $m$, let $g^m_{\text{inner},i} = \nabla_{\phi^m_0} \mathcal{L}^m_{\text{inner},i}(\phi^m_0)$. Assume that the inner objective is $\beta_m$-smooth along the update segment. For $0 < \eta_m < 2/\beta_m$, one patient-specific update satisfies*

$$\mathcal{L}^m_{\text{inner},i}(\phi^m_i) \leq \mathcal{L}^m_{\text{inner},i}(\phi^m_0) - \eta_m \left(1 - \frac{\beta_m \eta_m}{2}\right) \|g^m_{\text{inner},i}\|_2^2. \tag{10}$$

*Moreover, defining the historical residual $e^m_{ij} = u^m_{ij} - \mathcal{M}_{\phi^m_0}(k^m_{ij})$, the parameter Jacobian $J^m_0(x) = \partial \mathcal{M}_{\phi^m}(x)/\partial \phi^m|_{\phi^m = \phi^m_0}$, and the induced tangent kernel $\mathcal{K}^m_0(q,k) = J^m_0(q) J^m_0(k)^\top$, the adapted memory read for a current-image query $q_i$ satisfies*

$$\mathcal{M}_{\phi^m_i}(q_i) = \mathcal{M}_{\phi^m_0}(q_i) + \frac{2\eta_m}{N^m_i} \sum_{j=1}^{N^m_i} \mathcal{K}^m_0(q_i, k^m_{ij}) e^m_{ij} + \mathcal{R}^m_i, \tag{11}$$

*where $\|\mathcal{R}^m_i\|_2 = O(\eta_m^2 \|g^m_{\text{inner},i}\|_2^2)$ when the parameter Jacobian is locally Lipschitz. Let $\mathcal{H}^m_{\text{inner},i} = \nabla^2_{\phi^m_0} \mathcal{L}^m_{\text{inner},i}(\phi^m_0)$ denote the inner-loss Hessian, $\mathbf{I}$ the identity matrix on the memory parameter space, and $B$ the number of cases in the batch. The outer objective differentiates through the same update as*

$$\nabla_{\phi^m_0} \mathcal{L}_{\text{outer}} = \sum\nolimits_{i=1}^{B} \left(\mathbf{I} - \eta_m \mathcal{H}^m_{\text{inner},i}\right)^\top \nabla_{\phi^m_i} \mathcal{L}_{\text{outer}}, \tag{12}$$

*and therefore learns the parameters governing the induced retrieval mechanism. Specifically, let $h^{\text{cur}}$ denote a current visual feature and $h^m$ a historical feature from modality $m$, and let $\text{Norm}(\cdot)$ denote row-wise $\ell_2$ normalization. For the globally learned parameters $\Theta$, the induced feature-space kernel is*

$$\mathcal{K}^m_\Theta(h^{\text{cur}}, h^m) = J^m_0\Big(\text{Norm}(h^{\text{cur}} W^{\text{img}}_K)\Big) J^m_0(\text{Norm}(h^m W^m_K))^\top. \tag{13}$$

*If the modalities instead share fast weights, their joint update contains the additional first-order interaction $-\eta \langle g^{\text{img}}_{\text{inner},i}, g^{\text{txt}}_{\text{inner},i} \rangle$. Separate memories remove this cross-modal writing term.*

Theorem 1 shows that the inner update constructs a patient-specific residual correction, while the current-image query determines how historical residuals contribute to the retrieved prompt. Outer report supervision in turn learns the initialization and projections that determine this retrieval mechanism. The proof is provided in Appendix A.

**Theorem 2** (Context-advantage decomposition and score-space anchoring)**.** *Fix the learned representations and consider a reduced context $k \in \{T, V, \emptyset\}$ paired with contribution $a \in \{V, T, H\}$ through $(a,k) \in \{(V,T), (T,V), (H,\emptyset)\}$. Let $L_i$ denote the number of supervised tokens in the current report, $A_{a,i} = S^F_i - S^k_i$ the context advantage. The corresponding sequence-level advantage is*

$$\widetilde{A}_{k,i} = L_i A_{a,i} = L_i(S^F_i - S^k_i) = \log \frac{p_\theta(R^{\text{cur}}_i \mid \mathcal{Z}^F_i)}{p_\theta(R^{\text{cur}}_i \mid \mathcal{Z}^k_i)}. \tag{14}$$

*Assuming finite relevant expectations,*

$$\mathbb{E}[\widetilde{A}_k] = \text{MI}_k - D_F + D_k, \tag{15}$$

*where* $\mathrm{MI}_k = \mathrm{MI}(R^{\mathrm{cur}}; \mathcal{Z}^F \mid \mathcal{Z}^k)$ *is the conditional mutual information contributed by the full context beyond the reduced context, and*

$$D_c = \mathbb{E}_{\mathcal{Z}^c} D_{\mathrm{KL}}(p(\cdot \mid \mathcal{Z}^c) \,\|\, p_\theta(\cdot \mid \mathcal{Z}^c)), \qquad c \in \{F, k\}, \tag{16}$$

*denotes the corresponding expected conditional fitting error. Here,* $p$ *and* $p_\theta$ *denote the true and model report distributions, respectively. When both model conditionals match the true conditionals, the sequence-level expectation reduces to the nonnegative conditional mutual information. The common report-dependent length normalization also preserves nonnegativity of the expected normalized advantage. For the normalized advantages optimized by CECA, let* $\bar{A}_a = \frac{1}{B}\sum_{i=1}^{B} A_{a,i}$ *denote the batch-mean advantage and let* $\delta_a$ *be its calibrated margin. Convexity gives*

$$[\delta_a - \bar{A}_a]_+ \le \frac{1}{B}\sum\nolimits_{i=1}^{B} [\delta_a - A_{a,i}]_+. \tag{17}$$

*Furthermore, let* $\bar{S}^c = \frac{1}{B}\sum_{i=1}^{B} S_i^c$ *denote the batch-mean score under context* $c \in \{F, T, V, \emptyset\}$*,* $\lambda_a$ *the CECA weight for contribution* $a$*, and* $\chi_a = \mathbf{1}\{\bar{A}_a < \delta_a\}$ *the indicator of an active advantage constraint. If* $\alpha_T \ge \lambda_V$*,* $\alpha_V \ge \lambda_T$*, and* $\alpha_\emptyset \ge \lambda_H$*, then away from hinge boundaries,*

$$\frac{\partial \mathcal{L}_{\mathrm{outer}}}{\partial \bar{S}^F} = -1 - \sum\nolimits_{a \in \{V,T,H\}} \lambda_a \chi_a < 0, \qquad \frac{\partial \mathcal{L}_{\mathrm{outer}}}{\partial \bar{S}^k} = -\alpha_k + \lambda_a \chi_a \le 0. \tag{18}$$

*The same signs hold for admissible subgradients of the hinge function.*

Theorem 2 shows that context advantage measures the additional predictive information of full history together with differences in model fitting error, while the mean hinge allows case-level variation. Most importantly, under the stated weight conditions, decreasing a reduced-context score alone cannot improve the objective. Instead, active CECA constraints strengthen the incentive to improve full-context prediction. The proof is provided in Appendix B.

## 4 Experiments

**Dataset.** We conduct experiments on Longitudinal-MIMIC (Zhu et al., 2023), a benchmark derived from MIMIC-CXR (Johnson et al., 2019) for evaluating temporal modeling in radiology report generation. It contains 95,169 longitudinal samples from 26,625 patients. Following the partition used in recent longitudinal report generation studies (Liu et al., 2025c; Dong et al., 2026; Song et al., 2025), we use 92,374 training samples from 26,156 patients, 737 validation samples from 203 patients, and 2,058 test samples from 266 patients.

**Evaluation metrics.** We evaluate generated reports using both natural language generation (NLG) and clinical efficacy (CE) metrics. The NLG metrics include BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005) and ROUGE-L (Lin, 2004), which assess textual similarity between generated and reference reports. For clinical evaluation, we use CheXbert (Smit et al., 2020) to extract labels for 14 clinical observations from generated and reference reports and compute precision, recall, and F1-score.

**Implementation details.** We implement PHAM-CECA with Swin Transformer (Liu et al., 2021) as the visual encoder and LLaMA-2-7B-Chat (Touvron et al., 2023) as the report generator. The language model operates in FP16, while memory adaptation is performed in FP32. All stages are optimized using AdamW with a learning rate of $2 \times 10^{-5}$ and weight decay of 0.05. PHAM is initialized from the epoch-5 SFT checkpoint, and CECA from the validation-selected epoch-2 PHAM checkpoint. Each modality-specific memory uses a $768 \to 2048 \to 4096$ SiLU-activated multilayer perceptron. We use one differentiable inner update with $\eta_{\mathrm{img}} = \eta_{\mathrm{txt}} = 0.1$ and an $\ell_2$-normalization constant of $10^{-6}$. We set $\alpha_T = \alpha_V = \alpha_\emptyset = 0.1$, $\lambda_V = \lambda_T = \lambda_H = 0.05$, and average the three hinge terms. Margin calibration uses $\rho = 0.5$ and $\gamma = 0.05$.

### 4.1 Comparison with state-of-the-art methods.

We compare PHAM-CECA with representative longitudinal RRG methods, including Prefilling (Zhu et al., 2023), HERGen (Wang et al., 2024), HC-LLM (Liu et al., 2025c), and BiOTPrompt (Liu et al., 2026). As shown in Table 1, PHAM-CECA achieves strong performance across NLG and CE metrics.

Its clinical F1 reaches 0.483, compared with 0.474 for Diff-RRG (Yun et al., 2025), while its ROUGE-L exceeds BiOTPrompt (0.296 versus 0.285). The baseline variants further illustrate the benefits of historical information. Adding either prior images or reports improves the Current-only baseline, with larger gains from reports and further improvements when both modalities are available. However, simply incorporating complete history remains less effective than PHAM-CECA. Existing approaches exploit history through mechanisms such as HC-LLM's representation consistency constraints and BiOTPrompt's evolution-aware visual prompts. PHAM-CECA instead combines patient-specific memory adaptation with explicit optimization of historical contributions, connecting how history is retrieved with how it supports report prediction. Its additional F1 improvement over the Full-history baseline, from 0.477 to 0.483, supports the benefit of this joint design beyond historical input availability.

Table 1: Comparison with existing methods on Longitudinal MIMIC-CXR using natural language generation (NLG) and clinical efficacy (CE) metrics. CE is measured by precision (P), recall (R), and F1-score. The best results are bolded, and the second-best results are underlined.

| Method | Natural Language Generation | | | | | | Clinical Efficacy | | |
|---|---|---|---|---|---|---|---|---|---|
| | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGE-L | P | R | F1 |
| Prefilling Zhu et al. (2023) | 0.343 | 0.210 | 0.141 | 0.100 | 0.137 | 0.274 | 0.506 | 0.364 | 0.423 |
| HERGen Wang et al. (2024) | 0.389 | 0.242 | 0.163 | 0.117 | 0.155 | 0.282 | 0.421 | 0.289 | 0.295 |
| STREAM Yang et al. (2025) | 0.394 | 0.237 | 0.144 | 0.104 | 0.143 | 0.261 | 0.472 | 0.428 | 0.411 |
| LLM-RG4 Wang et al. (2025) | 0.417 | 0.240 | 0.155 | 0.115 | 0.147 | 0.257 | 0.498 | 0.441 | 0.436 |
| Diff-RRG Yun et al. (2025) | 0.405 | 0.251 | 0.169 | <u>0.120</u> | 0.164 | 0.276 | 0.528 | 0.430 | 0.474 |
| MLRG Liu et al. (2025b) | 0.416 | 0.252 | 0.157 | 0.114 | 0.158 | 0.264 | 0.507 | 0.425 | 0.418 |
| HC-LLM Liu et al. (2025c) | 0.404 | 0.247 | 0.164 | 0.116 | 0.163 | 0.271 | 0.488 | 0.415 | 0.448 |
| BiOTPrompt Liu et al. (2026) | 0.397 | <u>0.253</u> | <u>0.174</u> | **0.126** | 0.155 | 0.285 | 0.471 | 0.424 | 0.417 |
| Baseline(Curent-only) | 0.395 | 0.234 | 0.157 | 0.103 | 0.163 | 0.276 | 0.503 | 0.428 | 0.451 |
| +historical image | 0.401 | 0.240 | 0.162 | 0.107 | 0.168 | 0.283 | 0.511 | 0.437 | 0.459 |
| +historical report | 0.410 | 0.247 | 0.170 | 0.112 | 0.174 | 0.289 | 0.523 | 0.449 | 0.468 |
| +historical report & image | <u>0.418</u> | 0.252 | 0.173 | 0.116 | <u>0.178</u> | <u>0.293</u> | <u>0.531</u> | <u>0.460</u> | <u>0.477</u> |
| PHAM-CECA(Ours) | **0.424** | **0.258** | **0.178** | 0.118 | **0.181** | **0.296** | **0.537** | **0.466** | **0.483** |

### 4.2 Ablation Studies

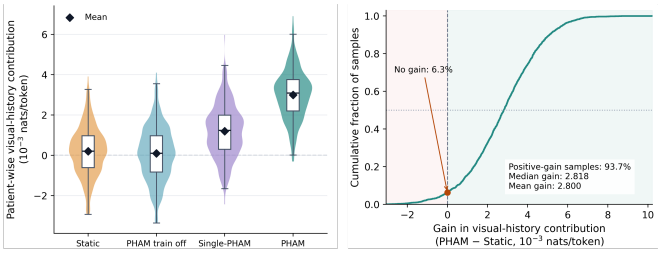
**Historical representation.** We compare three historical representation strategies to examine the benefits of patient-specific adaptation and current-conditioned retrieval. *Direct Concat.* directly combines historical and current features, while *Static* maps visual and textual history into current-independent prompts without patient-specific adaptation. *PHAM* instead adapts separate historical memories and queries them using the current image. Table 2 shows that *Static* improves NLG scores over *Direct Concat.*, suggesting that static historical encoding can improve textual similarity. *PHAM* achieves higher scores across both metric groups. These results support the effectiveness of patient-specific adaptation and current-conditioned historical retrieval in improving both textual and clinical agreement with reference reports.

Table 2: Comparison of historical representation strategies, including direct feature concatenation (Direct Concat.), static historical prompts (Static), and patient-specific associative memory (PHAM).

| Setting | Natural Language Generation | | | | | | Clinical Efficacy | | |
|---|---|---|---|---|---|---|---|---|---|
| | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGE-L | P | R | F1 |
| Direct Concat. | 0.365 | 0.216 | 0.149 | 0.084 | 0.158 | 0.254 | 0.493 | 0.419 | 0.441 |
| Static | 0.378 | 0.228 | 0.158 | 0.090 | 0.159 | 0.263 | 0.482 | 0.408 | 0.429 |
| PHAM | 0.392 | 0.237 | 0.165 | 0.096 | 0.172 | 0.274 | 0.516 | 0.443 | 0.465 |

**Visual-history contribution analysis.** We further examine whether PHAM improves visual-history contribution both on average and across individual cases. As shown in Figure 2, the contribution of history $A_V$ is distributed around zero in *Static* and non-adapative PHAM, whereas Single-PHAM shifts the distribution positively and full PHAM yields the largest increase. This trend supports the benefit of patient-specific adaptation and modality-specific memories. As reported in Table 3, PHAM

improves over Static by $2.805 \times 10^{-3}$ nats/token on average, with a 95% CI of $[2.697, 2.907] \times 10^{-3}$ and positive gains for 93.7% of samples. PHAM also exceeds Single-PHAM and non-adaptive PHAM by $1.801 \times 10^{-3}$ and $2.892 \times 10^{-3}$ nats/token, respectively. These results indicate that the improvement in visual-history contribution is broadly distributed across cases and is closely associated with patient-specific memory adaptation.




(a) Patient-wise visual-history contribution (b) Paired gains

Figure 2: Patient-wise visual-history contributions. (a) Contribution distributions across PHAM variants and (b) paired gains over Static.

Table 3: Paired gains in visual-history contribution. $\Delta A_V$ and the 95% confidence intervals are reported in units of $10^{-3}$ nats/token. Pos. denotes the percentage of samples with a positive paired gain.

| Comparison | $\Delta A_V$ | 95% CI | Pos. (%) |
|---|---|---|---|
| vs. Static | 2.805 | [2.697, 2.907] | 93.7 |
| vs. Single-PHAM | 1.801 | [1.693, 1.908] | 84.3 |
| vs. train off | 2.892 | [2.800, 2.996] | 95.5 |

**Context-advantage optimization.** We ablate reduced-context supervision and advantage constraints with the PHAM architecture fixed, while retaining full-context generation supervision for all variants. $A_V$, $A_T$, and $A_H$ measure the relative likelihood gains contributed by prior-image, prior-report, and complete-history information, respectively. As shown in Table 4, reduced-context supervision improves F1 score from 0.466 to 0.474. Advantage constraints alone increase $A_V$ and $A_H$ but lower F1 score, indicating that larger context advantages do not necessarily imply better clinical quality. Combining both components yields the best overall performance. These results demonstrate that reduced-context supervision and advantage optimization are complementary: the former preserves generation quality across contexts, while the latter explicitly strengthens the model's reliance on historical information. A complementary qualitative comparison across the four context settings is provided in Appendix D, where removing historical modalities progressively reduces the ability to describe both interval changes and persistent findings.

Table 4: Ablation of CECA with the PHAM architecture fixed. All variants retain full-context generation supervision. The two additional objective components are *Reduced CE*, defined as $\alpha_T \mathcal{L}_T + \alpha_V \mathcal{L}_V + \alpha_\emptyset \mathcal{L}_\emptyset$, and *Advantage*, defined as $\mathcal{L}_{\text{CECA}}$.

| Setting | Objective Components | | Clinical Efficacy | | | Context Advantage | | |
|---|---|---|---|---|---|---|---|---|
| | Reduced CE | Advantage | P | R | F1 | $A_V$ | $A_T$ | $A_H$ |
| Full-context supervision | × | × | 0.519 | 0.444 | 0.466 | 0.008 | 0.045 | 0.057 |
| w/o advantage constraints | ✓ | × | 0.528 | 0.454 | 0.474 | 0.009 | 0.046 | 0.048 |
| w/o reduced-context supervision | × | ✓ | 0.523 | 0.447 | 0.469 | 0.021 | 0.046 | 0.069 |
| PHAM-CECA | ✓ | ✓ | 0.537 | 0.466 | 0.483 | 0.030 | 0.061 | 0.075 |

### 4.3 Model Analysis

**History-sensitive representation analysis.** We examine whether patient-specific adaptation and contribution-aware optimization preserve historical information beyond the current finding state and support more accurate report content. Figure 3 provides a t-SNE (Van der Maaten & Hinton, 2008) illustration of historical representations for *pleural effusion* under image-history input, covering four longitudinal states: *Not detected*, *New*, *Persistent*, and *Resolved*. The key comparisons are New versus Persistent and Resolved versus Not detected, since each pair shares the current status but differs in its prior state. Static and PHAM (train off) exhibit substantial mixing, whereas PHAM shows clearer local groupings and PHAM-CECA displays a further reduction in overlap. These patterns suggest that patient-specific adaptation improves the retention of history-sensitive visual information, while contribution-aware optimization further strengthens the separation of longitudinal states.

**Qualitative report analysis.** We additionally compare reports generated from the same longitudinal input in Figure 4. The reference reports describe "clear lungs" previously and "moderate interstitial

edema" currently, alongside persistent "cardiac enlargement". Static generates generic normal findings and misses both current abnormalities. PHAM captures "interstitial prominence" and "cardiac enlargement", while PHAM-CECA more closely matches the "edema" description and severity, and additionally identifies "aortic tortuosity". This case shows that PHAM improves the use of historical information, while CECA further supports the generation of changed and persistent findings.

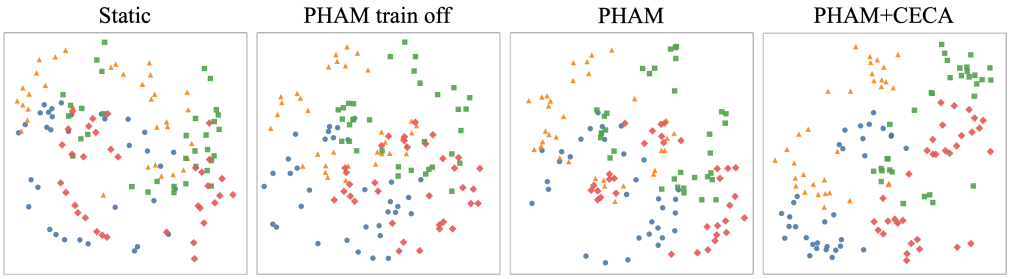
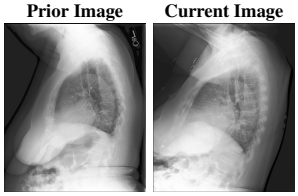



Figure 3: t-SNE visualization of longitudinal-state representations for *pleural effusion* under image-history input. Markers indicate ● *Not detected* ($0 \to 0$), ▲ *New* ($0 \to 1$), ■ *Persistent* ($1 \to 1$), and ◆ *Resolved* ($1 \to 0$), where 0 and 1 denote negative/unmentioned and positive report labels, respectively. PHAM (train off) disables patient-specific adaptation at the same PHAM checkpoint.




**Prior Report:** Heart size remains mildly enlarged with a left ventricular predominance. The aorta is unfolded and diffusely calcified, with the hilar contours appearing stable. The lungs are clear without evidence of pulmonary vascular engorgement. A trace left pleural effusion may be present, but no right pleural effusion is seen. No pneumothorax is identified. An inferior vena cava filter is noted within the abdomen. There are no acute osseous abnormalities.

**Current Report (Ground Truth):** Frontal and lateral views of the chest were obtained. There is prominence of the interstitial markings suggesting moderate interstitial edema. No large pleural effusion is seen. There is no evidence of pneumothorax. The cardiac silhouette is enlarged. The aorta is tortuous.

**Static (F1: 0.000, ROUGE-L: 0.261):** PA and lateral chest radiographs were obtained. The lungs are clear without focal airspace opacity. The heart size is within normal limits. No acute cardiopulmonary abnormality is identified.

**PHAM (F1: 0.667, ROUGE-L: 0.389):** Frontal and lateral chest radiographs were obtained. Mild diffuse interstitial prominence is present. The cardiac silhouette is enlarged. No pleural effusion or pneumothorax is identified. No focal airspace consolidation is seen.

**PHAM+CECA (F1: 0.800, ROUGE-L: 0.554):** Frontal and lateral chest radiographs demonstrate moderate interstitial pulmonary edema with diffuse prominence of the interstitial markings. No large pleural effusion or pneumothorax is identified. The cardiac silhouette is enlarged, and the thoracic aorta is tortuous. Mild bibasilar atelectatic change is present.

Figure 4: Qualitative comparison of longitudinal radiology report generation. Purple denotes temporally changing clinical findings, and green denotes clinical findings shared across prior and current reports. For generated reports, blue denotes clinically correct predictions consistent with the current ground truth, while orange marks contradictions or unsupported statements.

**Memory adaptation and retrieval.** We examine how patient-specific adaptation and current-conditioned retrieval support report prediction (Appendix C). A single inner update reduces reconstruction loss in both memories and changes the retrieved prompts. Current-image queries produce query-dependent omission sensitivity, while replacing them with fixed queries increases full-context NLL. Moreover, omitting highly sensitive historical associations causes larger NLL increases than random omission. These diagnostics support the connection between patient-specific adaptation, current-conditioned historical retrieval, and reference-report prediction.

## 5 Conclusion

We proposed PHAM-CECA for longitudinal radiology report generation, combining patient-specific historical representation with contribution-aware optimization. PHAM adapts separate visual and textual memories and retrieves historical evidence conditioned on the current image. CECA combines calibrated expected context-advantage constraints with generation supervision across four contexts to encourage the model to benefit from historical information. Our theoretical analysis connects patient-specific adaptation to current-conditioned retrieval and characterizes the contribution-aware objective. Experiments on Longitudinal-MIMIC show improved textual and clinical agreement, with ablations supporting the complementary benefits of adaptive historical representation and contribution-aware supervision.

## AI USE STATEMENT

Generative AI tools were used solely for language polishing, including improving grammar, clarity, readability, and academic writing style. All AI-assisted revisions were carefully reviewed and verified by the authors, who take full responsibility for the final content of the paper.

## ETHICS STATEMENT

This work uses publicly available and de-identified chest X-ray data for research purposes. The proposed longitudinal report generation framework is intended for research use and should not be considered a substitute for professional clinical assessment.

## REPRODUCIBILITY STATEMENT

We provide the model architecture, patient-specific memory adaptation, training objectives, hyperparameter settings, and evaluation protocols in the main paper and appendix. Additional implementation details are included in the supplementary material.

## REFERENCES

Oishi Banerjee, Hong-Yu Zhou, Subathra Adithan, Stephen Kwak, Kay Wu, and Pranav Rajpurkar. Direct preference optimization for suppressing hallucinated prior exams in radiology report generation. *arXiv preprint arXiv:2406.06496*, 2024.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, 2005.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *Advances in Neural Information Processing Systems*, 38:113506–113543, 2026.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1439–1449, 2020.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International joint Conference on Natural Language Processing (volume 1: long papers)*, pp. 5904–5914, 2021.

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

Yiheng Dong, Yi Lin, Shilong Huang, Xiyan Yang, and Xin Yang. TIM: Temporal decoupling with iterative mutual-refinement model for longitudinal radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6951–6961, 2026.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

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

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

Hong Liu, Dong Wei, Zhe Xu, Xian Wu, Yefeng Zheng, and Liansheng Wang. RRG-DPO: Direct preference optimization for clinically accurate radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 552–562. Springer, 2025a.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10348–10359. IEEE, 2025b.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5595–5603, 2025c.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13755–13765, 2026.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pp. 9992–10002. IEEE, 2021.

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, 2021.

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.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8237. IEEE, 2022.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, pp. 9355–9366. PMLR, 2021.

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*, pp. 1500–1519, 2020.

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.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): RNNs with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

Fuying Wang, Shenghui Du, and Lequan Yu. HERGen: Elevating radiology report generation with longitudinal data. In *European Conference on Computer Vision*, pp. 183–200. Springer, 2024.

Zhuhao Wang, Yihua Sun, Zihan Li, Xuan Yang, Fang Chen, and Hongen Liao. LLM-RG4: Flexible and factual radiology report generation across diverse input contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8250–8258, 2025.

Yan Yang, Xiaoxing You, Ke Zhang, Zhenqi Fu, Xianyun Wang, Jiajun Ding, Jiamei Sun, Zhou Yu, Qingming Huang, Weidong Han, et al. Spatio-temporal and retrieval-augmented modeling for chest X-ray report generation. *IEEE Transactions on Medical Imaging*, 44(7):2892–2905, 2025.

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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 152–161. Springer, 2025.

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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 189–198. Springer, 2023.

# A PROOFS FOR TASK-LEARNED ASSOCIATIVE RETRIEVAL

We follow the main-text notation. The index $m \in \{\mathrm{img}, \mathrm{txt}\}$ identifies a memory stream, $\phi^m \in \mathbb{R}^{p_m}$ contains its weights and biases, and the memory input and output dimension is $d_m = d_L$. Historical keys and values are held fixed when differentiating the inner objective with respect to $\phi^m$. Their dependence on the projections remains part of outer differentiation. The derivations assume differentiability along the relevant parameter paths, including normalization at its evaluated inputs.

The rowwise normalized projections are $K_i^m = \mathrm{Norm}(H_i^m W_K^m)$ and $U_i^m = \mathrm{Norm}(H_i^m W_U^m)$. Each memory $\mathcal{M}_{\phi^m} : \mathbb{R}^{d_L} \to \mathbb{R}^{d_L}$ uses the inner objective and update

$$\mathcal{L}_{\mathrm{inner},i}^m(\phi^m) = \frac{1}{N_i^m} \sum\nolimits_{j=1}^{N_i^m} \left\| \mathcal{M}_{\phi^m}(k_{ij}^m) - u_{ij}^m \right\|_2^2, \quad \phi_i^m = \phi_0^m - \eta_m \nabla_{\phi_0^m} \mathcal{L}_{\mathrm{inner},i}^m(\phi_0^m). \tag{19}$$

Here $Z_i^m = \mathcal{M}_{\phi_i^m}(Q_i)$, and $Q_i = \mathrm{Norm}(H_i^{\mathrm{cur}} W_K^{\mathrm{img}})$ also enters the decoder directly. The outer objective $\mathcal{L}_{\mathrm{outer}} = \mathcal{L}_{\mathrm{gen}} + \mathcal{L}_{\mathrm{CECA}}$ learns the shared parameters $\Theta$, including memory initializations and projections, while each inner update changes only its case-specific memory parameters.

## A.1 ONE-STEP ADAPTATION AND STREAM SEPARATION

**Historical reconstruction.** Let $g_{\mathrm{inner},i}^m = \nabla_{\phi_0^m} \mathcal{L}_{\mathrm{inner},i}^m(\phi_0^m)$, and assume that the inner objective has a $\beta_m$-Lipschitz parameter gradient on a neighborhood containing the update segment. Applying the smoothness inequality with $\Delta\phi = -\eta_m g_{\mathrm{inner},i}^m$ yields

$$\begin{aligned} \mathcal{L}_{\mathrm{inner},i}^m(\phi_i^m) - \mathcal{L}_{\mathrm{inner},i}^m(\phi_0^m) &\le -\eta_m \|g_{\mathrm{inner},i}^m\|_2^2 + \frac{\beta_m \eta_m^2}{2} \|g_{\mathrm{inner},i}^m\|_2^2 \\ &= -\eta_m \left(1 - \frac{\beta_m \eta_m}{2}\right) \|g_{\mathrm{inner},i}^m\|_2^2. \end{aligned} \tag{20}$$

For $0 < \eta_m < 2/\beta_m$, a nonzero inner gradient gives a strict decrease in historical reconstruction loss, whereas a zero gradient leaves the memory unchanged. This guarantee concerns the inner objective under the stated local smoothness and step-size conditions.

**Cross-modal interference in a shared write.** Consider a shared-memory alternative with parameter vector $\phi$ and reconstruction gradients $g_{\mathrm{inner},i}^{\mathrm{img}}$ and $g_{\mathrm{inner},i}^{\mathrm{txt}}$. For the joint update $\phi' = \phi - \eta(g_{\mathrm{inner},i}^{\mathrm{img}} + g_{\mathrm{inner},i}^{\mathrm{txt}})$, a locally bounded Hessian gives

$$\begin{aligned} \mathcal{L}_{\mathrm{inner},i}^{\mathrm{img}}(\phi') - \mathcal{L}_{\mathrm{inner},i}^{\mathrm{img}}(\phi) &= \langle g_{\mathrm{inner},i}^{\mathrm{img}}, \phi' - \phi \rangle + O(\|\phi' - \phi\|_2^2) \\ &= -\eta \|g_{\mathrm{inner},i}^{\mathrm{img}}\|_2^2 - \eta \langle g_{\mathrm{inner},i}^{\mathrm{img}}, g_{\mathrm{inner},i}^{\mathrm{txt}} \rangle \\ &\quad + O\Big(\eta^2 \|g_{\mathrm{inner},i}^{\mathrm{img}} + g_{\mathrm{inner},i}^{\mathrm{txt}}\|_2^2\Big). \end{aligned} \tag{21}$$

Exchanging the modality labels gives the text-stream expression. A negative cross-modal inner product offsets the first-order decrease and, if sufficiently negative, can make the leading loss change positive. Independent memories remove this cross-modal inner-update term because each inner loss depends only on its own fast weights. This separates the local writing paths rather than establishing universal superiority over shared memory: aligned gradients can be beneficial, and shared queries and report supervision still couple the streams during outer optimization.

## A.2 KERNELIZED READING AND ITS TASK-DEPENDENT LEARNING

**Residual retrieval induced by one step.** Define $J_\phi^m(x) = \partial \mathcal{M}_\phi(x)/\partial\phi \in \mathbb{R}^{d_L \times p_m}$, $J_0^m(x) = J_{\phi_0^m}^m(x)$, and the historical residual $e_{ij}^m = u_{ij}^m - \mathcal{M}_{\phi_0^m}(k_{ij}^m)$. Differentiating Equation 19 gives

$$g_{\mathrm{inner},i}^m = \frac{2}{N_i^m} \sum\nolimits_{j=1}^{N_i^m} J_0^m(k_{ij}^m)^\top \left(\mathcal{M}_{\phi_0^m}(k_{ij}^m) - u_{ij}^m\right) = -\frac{2}{N_i^m} \sum\nolimits_{j=1}^{N_i^m} J_0^m(k_{ij}^m)^\top e_{ij}^m. \tag{22}$$

Consequently, the parameter increment is

$$\Delta\phi_i^m = \phi_i^m - \phi_0^m = \frac{2\eta_m}{N_i^m} \sum\nolimits_{j=1}^{N_i^m} J_0^m(k_{ij}^m)^\top e_{ij}^m. \tag{23}$$

For a token $q_i$ of $Q_i$, held fixed during the parameter expansion, the fundamental theorem of calculus gives

$$\mathcal{M}_{\phi_i^m}(q_i) = \mathcal{M}_{\phi_0^m}(q_i) + J_0^m(q_i)\Delta\phi_i^m + \mathcal{R}_i^m. \tag{24}$$

Here $\mathcal{R}_i^m = \int_0^1 [J^m_{\phi_0^m + s\Delta\phi_i^m}(q_i) - J_0^m(q_i)]\Delta\phi_i^m \, \mathrm{d}s$. Substituting Equation 23 yields

$$\mathcal{M}_{\phi_i^m}(q_i) = \mathcal{M}_{\phi_0^m}(q_i) + \frac{2\eta_m}{N_i^m} \sum\nolimits_{j=1}^{N_i^m} \mathcal{K}_0^m(q_i, k_{ij}^m) e_{ij}^m + \mathcal{R}_i^m, \tag{25}$$

where $\mathcal{K}_0^m(q,k) = J_0^m(q)J_0^m(k)^\top \in \mathbb{R}^{d_L \times d_L}$. The first-order correction therefore combines historical residuals through a query-dependent, matrix-valued tangent kernel. Its entries need not be nonnegative or sum to one.

If the parameter Jacobian at $q_i$ is $C_{J,m}$-Lipschitz along the update segment in operator norm, the integral remainder satisfies

$$\|\mathcal{R}_i^m\|_2 \le \int_0^1 C_{J,m} s \|\Delta\phi_i^m\|_2^2 \, \mathrm{d}s = \frac{C_{J,m}}{2}\eta_m^2 \|g^m_{\mathrm{inner},i}\|_2^2. \tag{26}$$

The remainder vanishes when the memory is affine in the adapted parameters. For the two-layer SiLU memory, the kernel expression characterizes the local first-order effect of adaptation.

**Learning the kernel through the same computation.** For a fixed batch and fixed adaptation rates, holding independently parameterized quantities other than $\phi_0^m$ fixed gives

$$\frac{\partial \phi_i^m}{\partial \phi_0^m} = \mathbf{I} - \eta_m \mathcal{H}^m_{\mathrm{inner},i}, \qquad \mathcal{H}^m_{\mathrm{inner},i} = \nabla^2_{\phi_0^m} \mathcal{L}^m_{\mathrm{inner},i}(\phi_0^m), \tag{27}$$

Here $\mathbf{I} \in \mathbb{R}^{p_m \times p_m}$. Since the initialization affects the outer objective only through the adapted memories, the chain rule yields

$$\nabla_{\phi_0^m} \mathcal{L}_{\mathrm{outer}} = \sum\nolimits_{i=1}^{B} \left(\mathbf{I} - \eta_m \mathcal{H}^m_{\mathrm{inner},i}\right)^\top \nabla_{\phi_i^m} \mathcal{L}_{\mathrm{outer}}. \tag{28}$$

Score normalization is already included in each fast-weight gradient, so the sum introduces no additional averaging. The visual-memory gradient aggregates the full-history and image-history contexts, while the text-memory gradient aggregates the full-history and text-history contexts.

Key and value projections receive outer gradients through historical writing. The shared visual key projection $W_K^{\mathrm{img}}$ additionally receives gradients through querying both memories and through the direct use of $Q_i$ in the decoder. These paths account for projection reuse without introducing an independent query or output projection.

For a current visual feature $h^{\mathrm{cur}} \in \mathbb{R}^{d_V}$ and a historical feature $h^m$ in its modality's encoder space, the kernel in feature coordinates is

$$\mathcal{K}^m_\Theta(h^{\mathrm{cur}}, h^m) = J_0^m\Big(\mathrm{Norm}(h^{\mathrm{cur}} W_K^{\mathrm{img}})\Big) \, J_0^m(\mathrm{Norm}(h^m W_K^m))^\top. \tag{29}$$

The learned initialization determines the kernel Jacobians, the key projections determine their input coordinates, and the value projections shape the historical residuals. Thus report supervision learns the one-step retrieval mechanism through the memory and projection parameters.

## B Proofs for calibrated expected context advantage

For the information-theoretic analysis, representation parameters are fixed, $p$ is the true data distribution, and $p_\theta$ is the model report distribution. The complete supervised report $R^{\mathrm{cur}}$ and its length $L = L(R^{\mathrm{cur}}) > 0$ are identical across contexts. Required expectations are finite, and the model assigns positive probability on the support of the corresponding true conditionals. Common instructions are suppressed. A retained stream uses identical retrieved prompts across contexts, so every reduced context is a deterministic component of the full context.

Reduced-context indices $k \in \{T, V, \emptyset\}$ and contribution indices $a \in \{V, T, H\}$ are paired as $(a,k) \in \{(V,T), (T,V), (H,\emptyset)\}$. Thus $A_{a,i} = S_i^F - S_i^k$ preserves the methodology's contribution definitions.

### B.1 LIKELIHOOD ADVANTAGES AND LENGTH NORMALIZATION

**Sequence likelihood and conditional information.** The autoregressive chain rule and the teacher-forced score give the exact sequence log-likelihood ratio

$$\widetilde{A}_{k,i} = L_i(S_i^F - S_i^k) = L_i A_{a,i} = \log \frac{p_\theta(R_i^{\mathrm{cur}} \mid \mathcal{Z}_i^F)}{p_\theta(R_i^{\mathrm{cur}} \mid \mathcal{Z}_i^k)}. \tag{30}$$

The normalized advantage is this log-likelihood ratio divided by $L_i$. Define the expected fitting error under context $c \in \{F, T, V, 0\}$ as

$$D_c = \mathbb{E}_{\mathcal{Z}^c} D_{\mathrm{KL}}(p(\cdot \mid \mathcal{Z}^c) \,\|\, p_\theta(\cdot \mid \mathcal{Z}^c)), \tag{31}$$

where the distributions inside the divergence are over current reports. Let $\mathrm{MI}_k = \mathrm{MI}(R^{\mathrm{cur}}; \mathcal{Z}^F \mid \mathcal{Z}^k)$ denote conditional mutual information. Adding and subtracting the true conditional log-likelihoods yields

$$\mathbb{E}[\widetilde{A}_k] = \mathbb{E}\log \frac{p(R^{\mathrm{cur}} \mid \mathcal{Z}^F)}{p(R^{\mathrm{cur}} \mid \mathcal{Z}^k)} - \mathbb{E}\log \frac{p(R^{\mathrm{cur}} \mid \mathcal{Z}^F)}{p_\theta(R^{\mathrm{cur}} \mid \mathcal{Z}^F)} + \mathbb{E}\log \frac{p(R^{\mathrm{cur}} \mid \mathcal{Z}^k)}{p_\theta(R^{\mathrm{cur}} \mid \mathcal{Z}^k)} = \mathrm{MI}_k - D_F + D_k. \tag{32}$$

The first term is conditional mutual information because $\mathcal{Z}^k$ is determined by $\mathcal{Z}^F$. When the model matches both true branch conditionals, $D_F = D_k = 0$ and the expected sequence-level advantage equals the nonnegative quantity $\mathrm{MI}_k$. It is strictly positive if the true conditionals differ on a set of positive probability, although individual ratios can be negative. For a fitted model, both fitting-error terms remain and must be included in the interpretation of the advantage. These identities concern the retrieved representations, not all information in the raw examinations.

**Length normalization.** Division by the same positive length preserves a case's advantage sign and branch ordering but changes relative weights across cases. In general, $\mathbb{E}[A_a]$ is neither $\mathbb{E}[\widetilde{A}_k]/\mathbb{E}[L]$ nor ordinary conditional mutual information.

Under true branch conditionals, Bayes' rule gives

$$\log \frac{p(R^{\mathrm{cur}} \mid \mathcal{Z}^F)}{p(R^{\mathrm{cur}} \mid \mathcal{Z}^k)} = \log \frac{p(\mathcal{Z}^F \mid R^{\mathrm{cur}}, \mathcal{Z}^k)}{p(\mathcal{Z}^F \mid \mathcal{Z}^k)}. \tag{33}$$

For continuous representations, the density ratio is interpreted via Radon–Nikodym derivatives. Since $1/L(R^{\mathrm{cur}})$ is constant conditional on $(R^{\mathrm{cur}}, \mathcal{Z}^k)$, taking the remaining expectation over $\mathcal{Z}^F$ yields

$$\mathbb{E}[A_a] = \mathbb{E}_{R^{\mathrm{cur}}, \mathcal{Z}^k}\left[\frac{D_{\mathrm{KL}}\big(p(\mathcal{Z}^F \mid R^{\mathrm{cur}}, \mathcal{Z}^k) \,\|\, p(\mathcal{Z}^F \mid \mathcal{Z}^k)\big)}{L(R^{\mathrm{cur}})}\right] \geq 0. \tag{34}$$

Thus common report-dependent length normalization preserves nonnegativity of the expected advantage under the stated matching assumption, but not the ordinary mutual-information identity.

### B.2 EXPECTED HINGES AND SCORE-SPACE ANCHORING

**Constraints on mean contributions.** The function $h_a(x) = [\delta_a - x]_+$ is convex, so Jensen's inequality gives

$$\left[\delta_a - \frac{1}{B}\sum\nolimits_{i=1}^{B} A_{a,i}\right]_+ \leq \frac{1}{B}\sum\nolimits_{i=1}^{B} [\delta_a - A_{a,i}]_+. \tag{35}$$

The mean constraint permits small or negative case advantages to be offset by larger ones. If a random batch mean $\bar{A}_a$ has expectation $\mu_a$, Jensen's inequality also gives

$$[\delta_a - \mu_a]_+ \leq \mathbb{E}_{\mathcal{B}}[\delta_a - \bar{A}_a]_+. \tag{36}$$

The expected batch penalty therefore upper-bounds, but need not equal, the penalty at the population mean. Sampling fluctuations can activate a batch constraint even when its population mean exceeds the margin.

**Absence of a direct degradation incentive.** Let $\bar{S}^c = B^{-1}\sum_i S_i^c$. With the contribution-context pairing above, the outer objective is

$$\mathcal{L}_{\text{outer}} = -\bar{S}^F - \sum\nolimits_{k\in\{T,V,\emptyset\}} \alpha_k \bar{S}^k + \sum\nolimits_{(a,k)\in\{(V,T),(T,V),(H,\emptyset)\}} \lambda_a[\delta_a - \bar{S}^F + \bar{S}^k]_+. \quad (37)$$

Away from hinge boundaries, set $\chi_a = \mathbf{1}\{\bar{A}_a < \delta_a\}$. Direct differentiation gives

$$\frac{\partial \mathcal{L}_{\text{outer}}}{\partial \bar{S}^F} = -1 - \sum\nolimits_{a\in\{V,T,H\}} \lambda_a\chi_a < 0, \qquad \frac{\partial \mathcal{L}_{\text{outer}}}{\partial \bar{S}^k} = -\alpha_k + \lambda_a\chi_a \le 0, \quad (38)$$

The second inequality follows from $\alpha_T \ge \lambda_V$, $\alpha_V \ge \lambda_T$, and $\alpha_\emptyset \ge \lambda_H$. At a hinge boundary, every admissible $\chi_a \in [0,1]$ preserves these signs. Lowering a reduced-context score alone cannot decrease the objective, while a positive-weight active constraint strengthens the incentive to increase the full-context score.

Conversely, $\alpha_T < \lambda_V$ gives $\partial\mathcal{L}_{\text{outer}}/\partial\bar{S}^T > 0$ under an active visual constraint, directly rewarding a lower text-history score. The weight conditions remove this incentive when generation losses and advantages share the same target masking and case-level normalization. This is a score-space property. Shared-parameter optimization can still trade off performance across contexts. With fixed margins and weights, the CECA contribution to the parameter gradient is

$$\nabla_\Theta \mathcal{L}_{\text{CECA}} = -\sum\nolimits_{a\in\{V,T,H\}} \lambda_a\chi_a \nabla_\Theta \bar{A}_a. \quad (39)$$

Together with generation supervision, this gradient propagates through the retrieved prompts and differentiable adaptation in Equation 28.

**Interpretation of calibrated margins.** Fix a warm-up model independently of the calibration sample. Let $\mu_a$ be its population mean normalized advantage on the calibration distribution, and suppose $b_a = \text{LCB}_\gamma(\widehat{A}_a)$ satisfies

$$\Pr\{b_a \le \mu_a\} \ge 1-\gamma. \quad (40)$$

On this coverage event, if $b_a > 0$, the margin $\delta_a = \rho\max\{0, b_a\}$ with $0 < \rho \le 1$ satisfies

$$0 < \delta_a = \rho b_a \le b_a \le \mu_a. \quad (41)$$

If $b_a \le 0$, the rule gives $\delta_a = 0$ without establishing $\mu_a \ge 0$. Negative batch advantages remain penalized when $\lambda_a > 0$.

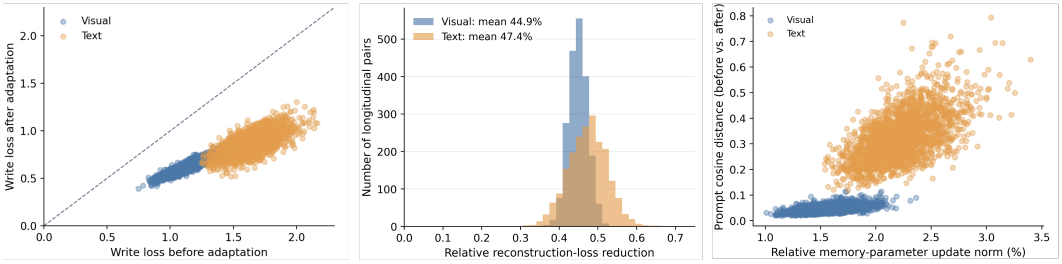
The coverage statement applies to each contribution for the fixed warm-up checkpoint on the calibration distribution and does not imply simultaneous guarantees across contributions. Once the margins are calibrated, they are treated as fixed optimization targets rather than renewed confidence guarantees for later checkpoints or other data distributions.

# C Additional analysis of PHAM

## C.1 Patient-specific adaptation dynamics

We examine whether the patient-specific inner update improves historical reconstruction and whether the resulting parameter changes affect the prompts retrieved by the current image. For each longitudinal pair and modality, we compare reconstruction losses before and after the single update in Equation 2. The relative reduction is measured as the loss decrease divided by the pre-update loss. We additionally examine relative memory-parameter update norms and the cosine distance between pre- and post-update prompts.

In Figure 5(a), the displayed before-after loss pairs lie below the identity line for both modalities, indicating that one inner update improves the reconstruction of historical associations. Figure 5(b) further shows that the reductions extend across longitudinal pairs, with mean relative decreases of 44.9% for visual memory and 47.4% for text memory. The textual stream exhibits a broader reduction distribution, reflecting greater variation in its response to adaptation. Figure 5(c) connects these updates to changes in the retrieved prompts. Larger relative parameter updates generally accompany larger prompt cosine distances, while textual prompts exhibit greater changes and variability than visual prompts. Thus, adaptation affects not only the reconstruction objective but also the representations supplied to the generator. These observations are consistent with the patient-specific correction described in Equation 11.

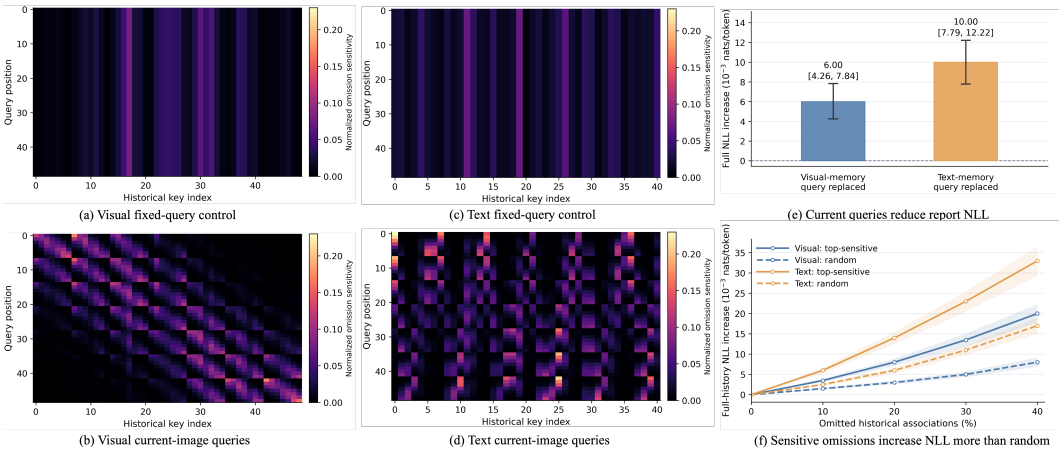



(a) One-step historical reconstruction (b) Distribution of reconstruction gains (c) Memory updates and prompt changes

Figure 5: Patient-specific adaptation dynamics. (a) Historical reconstruction loss before and after one inner update. The dashed line denotes equal losses. (b) Distributions of relative reconstruction-loss reductions. (c) Relative memory-parameter update norm versus the cosine distance between pre- and post-update prompts. Blue and orange denote visual and textual memories, respectively.

### C.2 Current-conditioned historical retrieval

We examine whether current-image queries produce query-dependent retrieval and whether the retrieved historical associations support reference-report prediction. We compare normalized omission-sensitivity maps under fixed and current-image queries for both memories. We further evaluate modality-specific query replacement and compare the omission of top-sensitive associations with random omission at matched rates. Prediction effects are measured by the increase in length-normalized full-history NLL relative to the unperturbed prediction of the same reference report. Positive changes indicate worse prediction.




(a) Visual fixed-query control (c) Text fixed-query control (e) Current queries reduce report NLL

(b) Visual current-image queries (d) Text current-image queries (f) Sensitive omissions increase NLL more than random

Figure 6: Current-conditioned historical retrieval. (a,c) Normalized omission sensitivity for visual and textual memories under fixed queries. (e) Mean full-history NLL increases after modality-specific query replacement. (b,d) Corresponding sensitivity maps under current-image queries. (f) Full-history NLL increases after omitting top-sensitive (solid) or randomly selected (dashed) historical associations.

**Query-dependent historical sensitivity.** Figures 6(a,c) show repeated sensitivity patterns across query positions under the fixed-query controls. In contrast, Figures 6(b,d) exhibit position-dependent patterns under current-image queries, indicating that different query positions are sensitive to different historical associations. This behavior is consistent with the query-dependent retrieval mechanism in Equation 11, where the current query modulates the contribution of historical residuals.

**Effects on report prediction.** Figure 6(e) shows that replacing visual-memory and text-memory queries increases mean full-history NLL by $6.00 \times 10^{-3}$ and $10.00 \times 10^{-3}$ nats/token, respectively. Figure 6(f) provides complementary evidence: omitting top-sensitive associations produces larger NLL increases than random omission at matched nonzero rates, with the gap widening as more associations are removed. At 40% omission, the plotted increases are approximately $20 \times 10^{-3}$ versus

$8 \times 10^{-3}$ nats/token for visual memory and $33 \times 10^{-3}$ versus $17 \times 10^{-3}$ nats/token for text memory. The difference from random omission suggests that the sensitivity ranking identifies associations with greater predictive relevance, rather than merely reflecting the amount of information removed. Together, these results connect current-conditioned retrieval to reference-report likelihood.

## D CASE STUDY UNDER DIFFERENT HISTORICAL-CONTEXT CONDITIONS.

Figure 7 provides a complementary qualitative analysis of how different historical contexts affect longitudinal report generation. Unlike the case study in the main text, which focuses on the improvement across model variants, this example keeps the prediction target fixed and compares the four context settings used throughout our evaluation: full history ($F$), text history only ($T$), visual history only ($V$), and current image only ($\emptyset$).

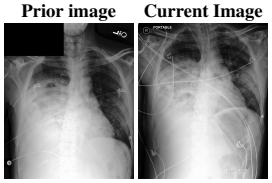
| | |
|---|---|
| **Prior image** **Current Image**  | **Prior Report:** Portable AP view of the chest demonstrates confluent opacity in the right mid and lower lung. There is relative sparing of the right apex. Rounded lucencies projecting over right hemithorax, are suggestive of cavities or abscess formation. Small-to-moderate right pleural effusion is likely. Ground-glass opacities most pronounced in the left mid lung zone. There is no large left pleural effusion. No pneumothorax is seen. Heart size is difficult to discern due to adjacent opacities. Partially imaged upper abdomen is unremarkable.<br>**Current Report (Ground Truth):** As compared to the previous radiograph, the patient has received a new orogastric tube. The course of the tube is unremarkable, the tip of the tube projects over the middle parts of the stomach. There is no evidence of complications. The other monitoring and support devices are in unchanged position. The massive bilateral lung abnormalities are constant in appearance. |
| **CECA-F: visual + text history (F1=1.000, ROUGE-L=0.692):** Compared with the previous radiograph, the patient has a new orogastric tube. The tube has an unremarkable course with its tip over the middle stomach. There is no evidence of complication. Other monitoring and support devices are unchanged. The massive bilateral lung abnormalities remain stable. | |
| **CECA-T: text history only (F1=0.667, ROUGE-L=0.457):** As compared to the previous radiograph, a new enteric tube has been placed with its tip projecting over the stomach. The other monitoring and support devices remain in unchanged position. Extensive bilateral pulmonary abnormalities show no significant interval change. A small right pleural effusion is present. | |
| **CECA-V: visual history only (F1=0.500, ROUGE-L=0.309):** As compared to the previous radiograph, bilateral lung abnormalities are unchanged. The monitoring and support devices remain in position. An enteric tube projects over the upper abdomen. A small right pleural effusion and mild bibasilar atelectasis are present. | |
| **CECA-$\emptyset$: current image only (F1=0.000, ROUGE-L=0.189):** There is bilateral lung abnormality. There is no evidence of pneumothorax. The heart is enlarged. | |

Figure 7: Qualitative case study under four historical-context settings: full history ($F$), text history only ($T$), visual history only ($V$), and current image only ($\emptyset$). Purple highlights temporally changing clinical findings in the current ground-truth report, while green denotes clinical findings shared by the prior and current reports. For generated reports, blue indicates clinically correct predictions consistent with the current ground truth, whereas orange indicates incorrect or unsupported clinical findings.

The current examination introduces a clinically relevant temporal change with the placement of a new orogastric tube, while the extensive bilateral pulmonary abnormalities remain largely stable from the prior study. With full historical context, CECA-F correctly captures both the newly introduced support device and the unchanged pulmonary findings, achieving the highest clinical F1 and ROUGE-L scores (1.000 and 0.692, respectively). Removing the prior image in CECA-T preserves most of the temporal information available from the previous report, but introduces an unsupported pleural-effusion finding, reducing the scores to 0.667 and 0.457. CECA-V retains coarse information about the persistent bilateral abnormalities and support devices but provides less precise temporal characterization and introduces additional false findings, resulting in further degradation (F1=0.500, ROUGE-L=0.309). In contrast, the current-only setting fails to recover the key historical change in support devices and produces only a generic description of the pulmonary abnormality, yielding the lowest performance (F1=0.000, ROUGE-L=0.189). The progressive degradation from $F$ to $T$, $V$, and $\emptyset$ qualitatively illustrates that the two historical modalities provide complementary information: the prior report offers explicit longitudinal semantics, whereas the prior image supplies additional visual evidence for distinguishing persistent findings from interval changes.