FROM EXPERIENCE TO LESSONS: ORGANIZING VISUAL MEMORY FOR CHEST X-RAY REPORT GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents have been found promising for completing complex tasks, where the episodes recording how an agent made decisions to obtain the outcomes can be maintained in a memory store for experience reuse to make better decisions. In this paper, we investigate the use of MLLM agents for chest X-ray report generation. We first make use of a multimodal (vision-language) episodic memory store for maintaining the experience. We further propose to organize the episodes as a compact set of multimodal lessons which are expected to be more generalizable for reuse. We conduct experiments on MIMIC-CXR with 76 lessons learned for retrieval by the agent to enhance the report generation quality, resulting in improvement on finding-level balanced accuracy by 2.90 points, more than retrieval over all 8,192 raw episodes (1.07 points). In addition, by allowing the agent to learn how to properly incorporate the retrieved lessons via rewarding corrected findings and penalizing those introducing errors, the gain over the agent without memory rises to 5.18 points and the introduced errors fall from 123 to 74. We also demonstrate that the lessons learned are transferable to different X-ray datasets, different agent frameworks, and different MLLM base models.

1 INTRODUCTION

Agents can improve by learning from their past experience, for instance, by storing verbal reflections on feedback and using them to guide later decisions (Shinn et al., 2023). For agents that work with images, how this experience is stored matters. As the raw episodes contain images, the growth in the size of the memory store is an issue. Abstracting experience into text makes it compact, but text summaries omit visual details needed to revisit an interpretation (Guo et al., 2026). Also, the raw episodes are sometimes too specific to the data they came from, limiting the experience reuse. We study this problem in the context of chest X-ray report generation. A reporting mistake could come from textual and visual modalities: what was wrongly said, and where it was wrongly located. A useful memory of past experience must keep both, while staying small enough to deploy and general enough to apply to new cases (Figure 1).

Existing work offers several routes for learning from experience. Agents can extract reusable insights and workflows from past trajectories (Zhao et al., 2024; Wang et al., 2025b), learn to use retrieved memories (Yan et al., 2026), or evolve skills derived from experience alongside a policy (Xia et al., 2026). Multimodal memory keeps source images together with guidance about perception and reasoning (Bo et al., 2026), and experience memory for medical diagnosis organizes past cases around confusable diagnoses (Feng et al., 2026). We ask whether an agent can turn its experience into compact knowledge that keeps its image evidence and still can generalize well to new data, different agent frameworks, and different base models.

We propose a vision-language memory framework that consolidates an agent's experience to form "lessons", which consist of both the X-ray images and the textual details about the abnormalities. An *episodic store* records individual episodes, each linking a case's images to the agent's work traces, reports, and feedback. In addition, a *consolidated store* contains *lessons*: reusable guidance synthesized from related episodes. Each lesson connects an explanation of a recurring reporting error to marked image regions and states when its suggested change applies. A lesson therefore

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.

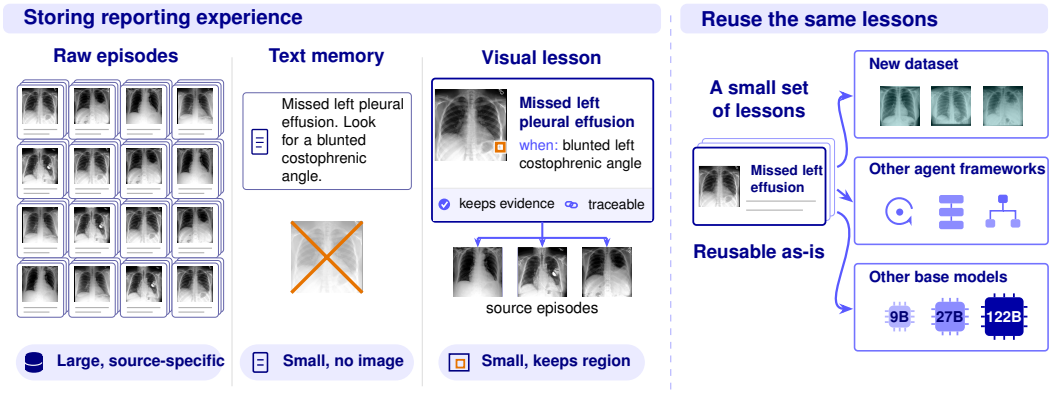


Figure 1: **Organizing visual experience into lessons.** Left: Three ways to store a reporting experience. Raw episodes keep images but form a large, source-specific store; text only memory is small but fails to retain visual information; a visual "lesson" is small, which keeps the marked region of abnormality with links to its source episodes. Right: The same lessons can be reused without modifications on new datasets, in other agent frameworks, and with other base models.

keeps provenance of a mistake: what was wrongly said and where it was wrongly located. Each lesson also records the episodes corresponding to it, so every change to a final X-ray report can be traced back to the underlying evidences. The episodic store is needed only to construct lessons. For deployment, only the consolidated store is needed, which is far smaller.

Applying a lesson to a new case requires judging whether the change it suggests fits that case. We further train the agent to make this judgment. The agent must interpret the lesson's explanation and conditions (*textual reasoning*) and check the suggested change against the current image (*visual grounding*). Training includes cases that need the change and cases where the same change would introduce errors, so the agent learns when a lesson applies and when it does not. We optimize the two judgments together through a report-level reward that credits corrected findings and penalizes introduced errors.

Experiments on MIMIC-CXR and IU-Xray show that the proposed visual-language memory framework can enhance the accuracy of an agent for X-ray report generation using only a compact set of 76 lessons automatically inferred from the raw episodes, with visual evidence effectively incorporated. For MIMIC-CXR, an agent with relevant lessons retrieved can raise finding-level balanced accuracy (FBA) by 2.90 points, while retrieval over all 8,192 source episodes raises it by only 1.07. For IU-Xray, adopting the same set of lessons raises FBA by 2.39 points without modification, while leveraging episodic memory gives no gain. We further show that the lessons learned using one agent framework can also improve the performance of three other agent frameworks, and for base models with 4B to 122B parameters. An agent built on a 4B model with lessons outperforms one built on a 27B model without them on both datasets. Removing or replacing the lesson images reduces the gains on both datasets. Finally, with the same set of lessons, training raises the number of corrected findings from 225 to 344 and lowers the introduced errors from 123 to 74.

Our contributions are as follows:

- **Lightweight and auditable visual-language memory.** We consolidate an agent's episodes into lessons that keep both textual and image evidence behind each X-ray reporting error. Deployment needs only the consolidated store, and every X-ray report change can be traced to its source episodes (Section 4.7).
- **Generalizable lessons.** Without modification, lessons improve the report generation on a new dataset, in other agent frameworks, and for base models with 4B to 122B parameters.
- **Optimizing how lessons are used.** We optimize textual reasoning and visual grounding together with a report-level reward, training on cases where the reference report supports a lesson's suggested change and cases where the report does not. This increases corrected findings and reduces introduced errors.

2 Related Work

Learning from agent experience. Agents can reuse past experience as verbal reflections (Shinn et al., 2023), extracted insights (Zhao et al., 2024), reusable step-by-step workflows (Wang et al., 2025b), or retrieved past cases (Zhou et al., 2025). Abstracting trajectories into text keeps memory compact, and such workflows can transfer across tasks, websites, and domains (Wang et al., 2025b). For image tasks, however, text summaries omit visual details needed to revisit an interpretation (Guo et al., 2026). Our lessons are abstracted in a similar way but keep the marked image regions and links to their corresponding episodes.

Visual memory and memory transfer. Several works keep visual or domain-specific experience in memory. AtlasVA combines spatial maps, visual exemplars, and symbolic skills for visual agents (Wang et al., 2026), and MedExpMem organizes diagnostic experience around confusable diagnoses (Feng et al., 2026). Closest to our work, ViLoMem stores visual error guidelines separately from logical ones, each written from a single error and kept with its whole source image (Bo et al., 2026). Each of our lessons is written from several errors of the same type and tested on held-out cases of that type; it marks the supporting region on each image and states when not to apply its change. MemCollab finds that memory built by one agent can hurt agents on other base models, and builds shared memory from several agents (Chang et al., 2026). Our lessons are written from a single agent's errors but describe image findings and their visual cues rather than the agent's reasoning, and they transfer unchanged to other base models, other frameworks, and a new dataset.

Learning to use memory. Memory-R1 trains an agent to select and reason over retrieved memories with answer rewards (Yan et al., 2026). MemSkill learns and evolves memory-operation skills (Zhang et al., 2026a), and SkillRL evolves a skill library during reinforcement learning (Xia et al., 2026). These methods train how memory is managed together with the policy. We focus on the step where a retrieved lesson is applied: with the store and retrieval fixed, we train the agent on cases where following a lesson corrects the draft and cases where it introduces an error, so it learns when not to apply a lesson.

Radiology report generation. Report generators, such as MAIRA-2 (Bannur et al., 2024) and CURV (Wang et al., 2025a), and medical multimodal models such as HuatuoGPT-Vision (Chen et al., 2024) and CheXOne (Zhang et al., 2026b), improve reporting by training on large collections of images and text. Retrieval-based methods find reports of similar past studies and use them as the new report or as input to a generator (Endo et al., 2021; Jeong et al., 2024; Sun et al., 2025), and agentic pipelines add drafting and verification stages (Abdaoui et al., 2026). Like our episodic store, retrieval-based methods draw on individual past cases. Our lessons are instead consolidated from past errors and state when they apply, and can be used by an existing agent without retraining.

3 Method

3.1 Task and framework overview

Report generation with lessons. Given chest X-ray images and available clinical context x, an agent first writes an initial draft d by following its own workflow, for example by viewing the radiograph and zooming into regions it selects. No memory is used at this stage. We record its work trace τ, let $o(\tau)$ be the pre-report observations, and define $\xi = (x, \tau, d)$. The agent then retrieves up to K lessons from a lesson store S and tries them as the context on the draft to produce the final report:

$$\hat{y} = \mathcal{R}_\theta(\xi; S), \tag{1}$$

where $\mathcal{R}_\theta$ denotes the full procedure of Section 3.3 (retrieval, lesson-conditioned revision and acceptance). We call the model that revises the draft with a lesson the reporting policy π_θ, which is the only component we train.

Two stores and three stages. An episodic store E records past episodes, each with its images, work trace, draft, reference report, and feedback. A consolidated store S, which we call the lesson store, holds lessons built from related errors in E. Figure 2 shows the three stages: constructing lessons from E (Section 3.2), generating a report with retrieved lessons (Section 3.3), and training π_θ to use them (Section 3.4). Reference reports y^* provide feedback during construction and rewards during training, but never enter the inputs used to report on a current case.

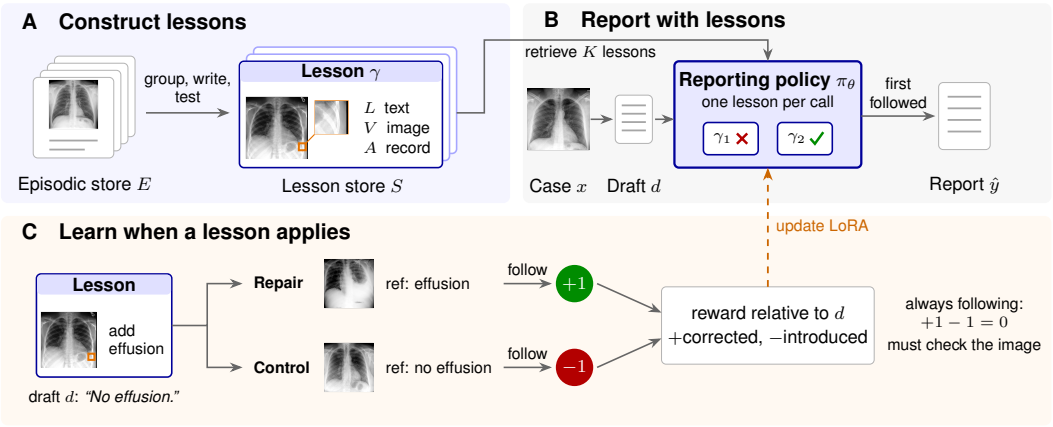


Figure 2: **From past experience to report generation.** (A) Related errors are consolidated into lessons with text L, images V, and record A. (B) Retrieved lessons are tried one at a time; the first revision that follows its lesson becomes the report. (C) Training pairs each lesson with repair cases (following it corrects the draft) and control cases (following it introduces an error).

Applying lessons to a draft. Lessons interact with an agent only through its draft and recorded context. Any agent that writes a draft can therefore use the same lesson store, whatever its actions or execution order. We use this property to apply one store to several agent frameworks (Section 4.6).

3.2 Consolidating experience into visual lessons

We consolidate recurring reporting errors into visual lessons, abstracting guidance across cases while keeping the image evidence behind it. We denote the construction procedure below (grouping, writing and testing) by C, so that $S = C(E)$.

Visual lessons. Each lesson consists of

$$\gamma = (L_\gamma, V_\gamma, A_\gamma). \tag{2}$$

The text L_γ has four fields: the visible condition that triggers the lesson, the reporting decision, the visual cue that separates presence from absence, and when not to apply it. The lesson images $V_\gamma = \{(I_{\gamma,j}, m_{\gamma,j})\}_{j=1}^{n_\gamma}$, $n_\gamma \in \{1,2,3\}$ pairs each of one to three historical images $I_{\gamma,j}$ with a marked region $m_{\gamma,j}$. The record A_γ stores the lesson's writing cases and its outcomes on its check cases. The agent receives L_γ and V_γ; A_γ is kept for inspection.

Grouping errors. Each episode $e = (x, \tau, d, y^*, \Phi)$ in E holds the images, the agent's work trace and draft, the reference report, and reference-based feedback Φ. We group the agent's errors by finding and the suggested update (Appendix A). Each group, and its lesson, has one target finding and one suggested update: *add*, when the reference reports the finding and the draft omits or denies it, or *remove*, when the draft asserts or hedges the finding and the reference rules it out. Each group is split into writing cases B_γ, used to write the lesson, and check cases H_γ, used to check it. The two sets come from different patients, and neither includes patients from the evaluation sets.

Writing a lesson. A writer reads the radiographs and reference reports of the writing cases and writes the lesson's condition, cue, and exception; the decision follows from the suggested update for the correction. The lesson keeps one to three frontal radiographs from writing cases whose reference supports the correction. On each image, the region the writer selected as visible support is outlined by its Chest ImaGenome box (Wu et al., 2021). The writer is asked to describe in the cue what these images show, so that the lesson text refers to the evidence the lesson images contain.

Testing and traceability. Before training, the reporting policy applies each lesson to the drafts of its check cases H_γ, and we record the findings it corrects and the errors it introduces. These outcomes are stored in A_γ with the source links.

3.3 MEMORY-GUIDED REPORT GENERATION

Retrieval selects lessons relevant to the current case, and the agent decides how they should affect its report.

Retrieval. Let $h(\xi)$ be a fixed embedding of the current case: its images, indication, and draft (Appendix B.5). We represent each lesson by the mean embedding of its writing cases, $k_\gamma = |B_\gamma|^{-1} \sum_{i \in B_\gamma} h(\xi_i)$, and rank lessons by the cosine similarity of their keys to $h(\xi)$, giving an ordered list $\Gamma(\xi)$ of up to K lessons. Matching the current case to the cases behind a lesson, rather than to the lesson's text, retrieves lessons by the situation in which the error occurred.

Lesson-conditioned revision. Lessons are tried one at a time, in retrieval order. Each attempt is a separate call that contains the current case, its recorded context, the draft, and exactly one lesson:

$$s_\gamma = (x, o(\tau), d, L_\gamma, V_\gamma), \qquad z_\gamma \sim \pi_\theta(\cdot \mid s_\gamma). \tag{3}$$

The policy is asked to use the current images to judge whether the lesson's suggested change fits the case. A proposal follows the lesson if the extracted label of its target finding changes in the lesson's suggested update (Section 3.4); this check uses the draft and the proposal, never the reference. One attempt returns $\mathcal{P}_\gamma(d, z_\gamma)$, which is z_γ if the proposal follows the lesson and d otherwise. The first attempt that returns a proposal becomes the final report, and the draft is kept if none does. Because each call contains one lesson, every change in the final report can be attributed to a single lesson and checked by removing it (Appendix D). Deciding whether to follow a lesson also reduces to a single-step judgment on the current case, which we optimize in the next section.

3.4 OPTIMIZING HOW LESSONS ARE USED

We train π_θ to make this judgment, with the lesson store and retrieval fixed. If training used only cases where a lesson's suggested change is correct, the policy could learn to follow every lesson it receives. We therefore pair each lesson with cases where following it corrects the draft and cases where it introduces an error, and score each proposal by how it changes the draft's findings.

Repair and control cases. Each training prompt pairs a case and its draft with one lesson, matching how lessons are tried at inference. Each prompt takes one of two roles, balanced for every lesson. In a *repair* case, the lesson's suggested change would correct a finding in the draft according to the reference report. In a *control* case, the reference supports the draft as it is, and the same change would introduce an error. Roles are assigned from reference reports and saved drafts; references never enter the policy's inputs.

For example, a lesson that adds effusion can meet the same draft statement, that effusion is absent, in both roles. Following it is correct when the reference reports effusion and harmful when it does not. A policy that always follows the lesson therefore gains on repair cases and loses on control cases. To earn reward in both roles, the policy must interpret the lesson's explanation and conditions (*textual reasoning*) and check them against the current images (*visual grounding*). A single report-level reward trains both, without separate reasoning annotations or localization targets.

Reward. For each graded finding $f \in \mathcal{F}(y^*)$, let $t_f = 1$ when the reference states that the finding is certainly or possibly present, and zero otherwise. We extract finding labels from a report y, setting $p_f(y) = 1$ for asserted or hedged findings and zero for denied, partly denied, or unmentioned findings. With $\tilde{y} = \mathcal{P}_\gamma(d, z)$ the report returned from an attempt, the reward measures the change in correct finding decisions relative to the draft:

$$r(z; s_\gamma, y^*) = \sum_{f \in \mathcal{F}(y^*)} \Big[\mathbf{1}\{p_f(\tilde{y}) = t_f\} - \mathbf{1}\{p_f(d) = t_f\} \Big]. \tag{4}$$

Each corrected finding earns $+1$ and each introduced error incurs -1, including changes to findings the lesson does not mention, so errors introduced on other findings offset gains on the lesson's finding. Malformed outputs receive -1.

Policy update. We maximize the expected reward with a GRPO-style objective (Shao et al., 2024). For each prompt, we sample G proposals, score their returned reports with Eqn. 4, and compute advantages by subtracting the group mean and divided by the standard deviation, so that no value

model is needed. Importance ratios are computed at the sequence level, as in GSPO (Zheng et al., 2025); the loss is summed over tokens and divided by a fixed constant, as in Dr. GRPO (Liu et al., 2025). A KL penalty toward the initial policy is applied. Only the language-side LoRA adapters (Hu et al., 2022) of π_θ are updated; the lesson store and retrieval stay fixed.

4 EXPERIMENTS

We organize the experiments around five questions. **RQ1:** Do lessons improve report generation? **RQ2:** Do lessons generalize to a different data distribution? **RQ3:** Does visual evidence in lessons help? **RQ4:** Does policy training improve how the agent uses lessons? **RQ5:** Do lessons generalize to other agent frameworks and model sizes?

4.1 EXPERIMENTAL SETUP

Data and models. We build lessons and train the policy on MIMIC-CXR (Johnson et al., 2019), using 8,192 construction studies and patient-disjoint validation and test sets of 1,000 studies each. To test generalization to a different data distribution, we also evaluate on 1,000 IU-Xray studies (Demner-Fushman et al., 2015), which never enter construction, training, or model selection. Unless stated otherwise, the agent is a ReAct agent (Yao et al., 2023) built on Qwen3.5-4B (Qwen Team, 2026), which also writes the lessons. The store holds 76 lessons (Appendix A), and retrieval returns up to $K = 3$. Appendix B gives the full setup, including training.

Compared systems. We compare with four existing report generation systems: Qwen3.5-4B generating the report directly, HuatuoGPT-Vision-7B (Chen et al., 2024), MAIRA-2 (Bannur et al., 2024), and CURV (Wang et al., 2025a). All use the same inputs and evaluation as our agent. Within our framework, we compare the agent without memory, with episodic memory, with ExpeL (Zhao et al., 2024) and ViLoMem (Bo et al., 2026), and with lessons; the two existing methods are built from the same writing cases and writer as the lesson store (Appendix B.2). Episodic memory retrieves the $K = 3$ most similar of the 8,192 episodes and tries each on the draft with the same policy, prompt structure, and acceptance rule as a lesson, so the two conditions differ only in what is retrieved. To test generalization, we apply the same store without modification to ReWOO (Xu et al., 2023), Plan-and-Solve (Wang et al., 2023), and a reimplementation of Swin-Qwen3 (Abdaoui et al., 2026), and to the ReAct agent on Qwen3.5-9B, 27B and 122B.

Metrics. We report BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee & Lavie, 2005) as lexical metrics, and RadGraph F1 (Jain et al., 2021), CheXbert F1 (Smit et al., 2020), GritLM similarity (Muennighoff et al., 2025), and FBA as clinical and semantic metrics. FBA is balanced accuracy over pooled finding-level decisions; its reference labels come from CAD-Chest (Zhang et al., 2023), appendix B.7 gives the definition of FBA. We report 95% case-level percentile bootstrap intervals from 5,000 draws shared across methods and metrics (Appendix B.3); paired differences use the same draws (Appendix C). All scores are multiplied by 100, and higher is better. Unless a caption states otherwise, subscripts give the half-width of the 95% interval, and within each dataset bold marks the best result and underline the second best.

4.2 LESSONS IMPROVE REPORT GENERATION (RQ1)

Lessons improve the agent's FBA on MIMIC-CXR from 64.79 to 67.69 (Table 1). With lessons, the agent's clinical and semantic scores are comparable to or above those of MAIRA-2 and CURV, the strongest existing systems in our evaluation. Gains are largest on metrics that score whether findings are stated correctly (FBA +2.90, CheXbert +5.20), smaller on RadGraph (+1.80), which scores entities and relations, and smallest on GritLM (+1.32), which scores overall similarity. Lexical metrics change by less than 0.3 points, consistent with lessons editing finding statements while leaving most wording unchanged. MAIRA-2 and CURV lead on ROUGE-L and METEOR respectively, which reward matching reference wording and favor systems trained to generate reports; our agent is not trained this way.

Lessons also outperform three other forms of memory built from the same experience. Retrieval over all 8,192 source episodes raises FBA by 1.07 and CheXbert by 2.20 points and lowers RadGraph by 0.70. ExpeL, which distills the writing cases into 6 global rules, and ViLoMem, which keeps

Table 1: **Report generation on MIMIC-CXR and IU-Xray.** Agent rows use ReAct on Qwen3.5-4B. All memory methods are built on MIMIC-CXR and applied to IU-Xray unchanged.

	Clinical and semantic				Lexical		
Method	RadGraph	CheXbert	GritLM	FBA	BLEU-4	ROUGE-L	METEOR
MIMIC-CXR							
Qwen3.5-4B	15.65$_{\pm 0.57}$	46.92$_{\pm 1.91}$	63.79$_{\pm 0.39}$	62.90$_{\pm 1.23}$	4.19$_{\pm 0.26}$	27.22$_{\pm 0.48}$	20.56$_{\pm 0.48}$
HuatuoGPT-Vision-7B	9.89$_{\pm 0.40}$	31.68$_{\pm 1.87}$	50.26$_{\pm 0.34}$	57.52$_{\pm 1.19}$	1.84$_{\pm 0.10}$	21.07$_{\pm 0.31}$	18.81$_{\pm 0.29}$
MAIRA-2	17.09$_{\pm 0.63}$	51.03$_{\pm 1.45}$	64.45$_{\pm 0.44}$	66.47$_{\pm 1.07}$	3.95$_{\pm 0.28}$	**30.61**$_{\pm 0.58}$	18.54$_{\pm 0.45}$
CURV	18.29$_{\pm 0.69}$	52.60$_{\pm 4.50}$	64.48$_{\pm 0.40}$	65.70$_{\pm 1.12}$	3.90$_{\pm 0.21}$	26.78$_{\pm 0.48}$	**22.99**$_{\pm 0.38}$
Agent, no memory	16.76$_{\pm 0.63}$	48.59$_{\pm 1.79}$	63.82$_{\pm 0.39}$	64.79$_{\pm 1.30}$	4.06$_{\pm 0.25}$	27.58$_{\pm 0.47}$	20.96$_{\pm 0.46}$
+ ExpeL	15.90$_{\pm 0.54}$	49.65$_{\pm 1.90}$	63.90$_{\pm 0.39}$	65.41$_{\pm 1.37}$	4.30$_{\pm 0.24}$	27.86$_{\pm 0.46}$	22.02$_{\pm 0.43}$
+ ViLoMem	15.75$_{\pm 0.55}$	49.14$_{\pm 1.93}$	63.78$_{\pm 0.40}$	65.22$_{\pm 1.36}$	4.11$_{\pm 0.24}$	27.55$_{\pm 0.47}$	21.48$_{\pm 0.44}$
+ episodic memory	16.06$_{\pm 0.58}$	50.79$_{\pm 1.80}$	63.72$_{\pm 0.37}$	65.86$_{\pm 1.38}$	**4.42**$_{\pm 0.26}$	28.06$_{\pm 0.48}$	21.73$_{\pm 0.45}$
+ lessons	**18.56**$_{\pm 0.69}$	**53.79**$_{\pm 1.93}$	**65.14**$_{\pm 0.41}$	**67.69**$_{\pm 1.39}$	3.83$_{\pm 0.23}$	27.56$_{\pm 0.48}$	20.68$_{\pm 0.46}$
IU-Xray							
Qwen3.5-4B	17.60$_{\pm 0.68}$	36.68$_{\pm 2.93}$	68.15$_{\pm 0.55}$	57.11$_{\pm 1.36}$	**8.42**$_{\pm 0.42}$	32.32$_{\pm 0.70}$	32.05$_{\pm 0.82}$
HuatuoGPT-Vision-7B	13.59$_{\pm 0.49}$	14.96$_{\pm 2.51}$	57.65$_{\pm 0.45}$	52.34$_{\pm 0.84}$	2.33$_{\pm 0.11}$	23.21$_{\pm 0.38}$	29.18$_{\pm 0.51}$
MAIRA-2	22.11$_{\pm 0.74}$	39.15$_{\pm 2.16}$	69.41$_{\pm 0.48}$	58.54$_{\pm 1.30}$	6.47$_{\pm 0.39}$	32.64$_{\pm 0.61}$	28.91$_{\pm 0.69}$
CURV	22.95$_{\pm 0.85}$	40.21$_{\pm 4.14}$	**69.56**$_{\pm 0.54}$	59.48$_{\pm 1.05}$	7.47$_{\pm 0.38}$	32.39$_{\pm 0.66}$	29.96$_{\pm 0.49}$
Agent, no memory	20.57$_{\pm 0.69}$	39.03$_{\pm 2.80}$	67.66$_{\pm 0.54}$	57.24$_{\pm 1.35}$	8.07$_{\pm 0.41}$	33.07$_{\pm 0.68}$	35.82$_{\pm 0.86}$
+ ExpeL	19.66$_{\pm 0.69}$	39.25$_{\pm 2.89}$	67.36$_{\pm 0.51}$	57.72$_{\pm 1.33}$	8.01$_{\pm 0.41}$	32.74$_{\pm 0.64}$	**35.97**$_{\pm 0.84}$
+ ViLoMem	19.71$_{\pm 0.69}$	39.46$_{\pm 2.85}$	67.60$_{\pm 0.53}$	57.72$_{\pm 1.33}$	8.14$_{\pm 0.42}$	32.87$_{\pm 0.65}$	35.91$_{\pm 0.86}$
+ episodic memory	20.08$_{\pm 0.68}$	37.60$_{\pm 2.80}$	67.84$_{\pm 0.53}$	57.23$_{\pm 1.33}$	7.86$_{\pm 0.40}$	33.03$_{\pm 0.66}$	35.25$_{\pm 0.86}$
+ lessons	**22.96**$_{\pm 0.79}$	**41.50**$_{\pm 2.96}$	68.84$_{\pm 0.56}$	**59.63**$_{\pm 1.46}$	7.85$_{\pm 0.41}$	**33.37**$_{\pm 0.69}$	35.27$_{\pm 0.87}$

1,070 per-case guidelines and retrieves them by similarity, raise FBA by 0.62 and 0.43 points and lower RadGraph by 0.86 and 1.01. The 76 lessons exceed episodic memory by 1.83 FBA and 3.00 CheXbert points and the two existing methods by at least 2.28 FBA and 4.14 CheXbert points. Lessons are not free of errors. Relative to the reports without lessons, they correct 225 finding decisions and introduce 123 errors, which motivates training the agent to apply a lesson only where the current images support it (Section 4.5).

4.3 Lessons generalize to IU-Xray (RQ2)

We apply the MIMIC-CXR store to IU-Xray without modification: no lesson is rewritten, and no IU-Xray study enters the store or the retrieval keys. On this dataset from a different institution, lessons still improve the agent (Table 1): FBA rises by 2.39 points, CheXbert by 2.47, RadGraph by 2.39, and GritLM by 1.18, while lexical metrics change by less than 0.6 points. Compared with MIMIC-CXR, the FBA and CheXbert gains are smaller and the RadGraph gain is larger. With lessons, the agent's FBA, CheXbert, and RadGraph are comparable to or above those of existing systems; CURV has the highest GritLM (69.56).

Episodic memory gives an FBA of 57.23, essentially unchanged to the agent without memory, and lowers CheXbert by 1.43 points. ExpeL and ViLoMem each raise FBA by 0.48 points and lower RadGraph by about 0.9. Both edit the text of many reports (167 and 94 of the 1,000) but change a finding decision in only 20 and 22 studies, correcting 29 and 28 decisions and introducing 7 and 6 errors. The two FBA values coincide because the net change in pooled counts is the same for both (+25 true positives, +3 false positives), not because they make the same edits. One explanation is that a retrieved episode or a per-case guideline records how one MIMIC-CXR case was reported, and six global rules may be too coarse to say when a change applies. A lesson instead states a correction together with the visual cue and the condition under which it applies, which are not tied to any single source case.

4.4 Visual evidence in lessons helps (RQ3)

Both the text and the images of lessons contribute to the gains (Table 2). Removing the images in lessons lowers FBA by 1.48 points on MIMIC-CXR and 1.61 points on IU-Xray, with CheXbert decreases of 2.43 and 1.61 points. Removing the lesson text while keeping the lesson images lowers

Table 2: **Ablation of visual evidence in lessons.** *w/o visual evidence* removes the lesson images; *w/o lesson text* keeps only the lesson images and marked regions; *random images* replaces the lesson images with randomly sampled chest X-rays. Current images are always given. All rows use the base Qwen3.5-4B policy.

	Clinical and semantic				Lexical		
Method	RadGraph	CheXbert	GritLM	FBA	BLEU-4	ROUGE-L	METEOR
MIMIC-CXR							
No memory	16.76$_{\pm 0.63}$	48.59$_{\pm 1.79}$	63.82$_{\pm 0.39}$	64.79$_{\pm 1.30}$	**4.06**$_{\pm 0.25}$	**27.58**$_{\pm 0.47}$	**20.96**$_{\pm 0.46}$
Lessons	**18.56**$_{\pm 0.69}$	**53.79**$_{\pm 1.93}$	**65.14**$_{\pm 0.41}$	**67.69**$_{\pm 1.39}$	3.83$_{\pm 0.23}$	27.56$_{\pm 0.48}$	20.68$_{\pm 0.46}$
w/o visual evidence	17.09$_{\pm 0.62}$	51.36$_{\pm 1.86}$	64.30$_{\pm 0.40}$	66.21$_{\pm 1.34}$	3.73$_{\pm 0.24}$	27.35$_{\pm 0.47}$	20.13$_{\pm 0.45}$
w/o lesson text	16.07$_{\pm 0.59}$	49.90$_{\pm 1.94}$	63.82$_{\pm 0.39}$	65.50$_{\pm 1.32}$	3.73$_{\pm 0.24}$	27.20$_{\pm 0.47}$	20.19$_{\pm 0.43}$
random images	16.89$_{\pm 0.63}$	50.70$_{\pm 1.79}$	64.12$_{\pm 0.40}$	65.49$_{\pm 1.30}$	3.91$_{\pm 0.24}$	27.42$_{\pm 0.47}$	20.24$_{\pm 0.44}$
IU-Xray							
No memory	20.57$_{\pm 0.69}$	39.03$_{\pm 2.80}$	67.66$_{\pm 0.54}$	57.24$_{\pm 1.35}$	**8.07**$_{\pm 0.41}$	33.07$_{\pm 0.68}$	**35.82**$_{\pm 0.86}$
Lessons	**22.96**$_{\pm 0.79}$	**41.50**$_{\pm 2.96}$	**68.84**$_{\pm 0.56}$	**59.63**$_{\pm 1.46}$	7.85$_{\pm 0.41}$	**33.37**$_{\pm 0.69}$	35.27$_{\pm 0.87}$
w/o visual evidence	22.06$_{\pm 0.74}$	39.89$_{\pm 2.80}$	68.22$_{\pm 0.55}$	58.02$_{\pm 1.39}$	7.85$_{\pm 0.41}$	33.21$_{\pm 0.67}$	35.38$_{\pm 0.88}$
w/o lesson text	21.92$_{\pm 0.74}$	38.59$_{\pm 2.80}$	68.09$_{\pm 0.56}$	56.84$_{\pm 1.29}$	7.70$_{\pm 0.39}$	33.19$_{\pm 0.68}$	35.13$_{\pm 0.88}$
random images	21.80$_{\pm 0.72}$	38.70$_{\pm 2.71}$	68.10$_{\pm 0.55}$	57.67$_{\pm 1.36}$	7.81$_{\pm 0.41}$	33.21$_{\pm 0.67}$	34.99$_{\pm 0.87}$

Table 3: **Base and trained policies on MIMIC-CXR.** Same lesson store and retrieval in all rows.

	Clinical and semantic				Lexical		
Method	RadGraph	CheXbert	GritLM	FBA	BLEU-4	ROUGE-L	METEOR
Base policy, no lessons	16.76$_{\pm 0.63}$	48.59$_{\pm 1.79}$	63.82$_{\pm 0.39}$	64.79$_{\pm 1.30}$	4.06$_{\pm 0.25}$	27.58$_{\pm 0.47}$	20.96$_{\pm 0.46}$
Base policy + lessons	18.56$_{\pm 0.69}$	53.79$_{\pm 1.93}$	65.14$_{\pm 0.41}$	67.69$_{\pm 1.39}$	3.83$_{\pm 0.23}$	27.56$_{\pm 0.48}$	20.68$_{\pm 0.46}$
Trained policy, no lessons	15.77$_{\pm 0.56}$	48.13$_{\pm 1.80}$	63.58$_{\pm 0.39}$	64.25$_{\pm 1.30}$	4.00$_{\pm 0.24}$	27.40$_{\pm 0.46}$	21.11$_{\pm 0.44}$
Trained policy + lessons	**18.80**$_{\pm 0.65}$	**56.84**$_{\pm 1.70}$	**66.45**$_{\pm 0.39}$	**69.97**$_{\pm 1.38}$	4.34$_{\pm 0.23}$	**28.03**$_{\pm 0.43}$	22.61$_{\pm 0.45}$
w/o visual evidence	18.64$_{\pm 0.65}$	54.74$_{\pm 1.68}$	65.26$_{\pm 0.39}$	68.56$_{\pm 1.33}$	**4.38**$_{\pm 0.24}$	27.94$_{\pm 0.43}$	**22.78**$_{\pm 0.43}$

FBA further, to 65.50 on MIMIC-CXR, only 0.71 points above the agent without memory, and to 56.84 on IU-Xray, below the agent without memory (57.24).

The benefit depends on the content of the lesson images, not on their presence. Replacing them with randomly sampled chest X-rays gives lower FBA than removing them on both datasets (65.49 versus 66.21 on MIMIC-CXR, and 57.67 versus 58.02 on IU-Xray), and the same holds for RadGraph, CheXbert, and GritLM. An image not showing the lesson's cue is therefore worse than no image: it gives the policy nothing to check the suggested change against, while the lesson text still asks it to.

4.5 Training increases the benefit of the same lessons (RQ4)

With the same store and retrieval, training raises corrected findings from 225 to 344 and lowers introduced errors from 123 to 74, relative to the reports without lessons (Table 3). Over the base policy, FBA rises by 2.28, CheXbert by 3.05, GritLM by 1.31, and RadGraph by 0.24 points; the complete system gains 5.18 FBA and 8.25 CheXbert points over the agent without memory. Without the lessons, the trained policy scores slightly below the base policy (64.25 vs. 64.79 FBA, with lower CheXbert and RadGraph), so the benefit of lessons grows from 2.90 to 5.72 FBA points. Removing the lesson images lowers the trained policy's FBA by 1.41 points, close to the 1.48-point drop for the base policy, so training does not shift the policy toward lesson text alone.

4.6 Lessons generalize across agent frameworks and model sizes (RQ5)

Agent frameworks. The lesson store was constructed from errors of the ReAct agent. Applied without modification, it raises FBA in all four frameworks on both datasets (Table 4), by 2.90 to 6.50 points on MIMIC-CXR and 1.69 to 3.78 points on IU-Xray. CheXbert and RadGraph also improve in all eight comparisons, although the RadGraph gains for ReWOO are small (0.22 and 0.07 points). On both datasets, the largest gain occurs for Plan-and-Solve, the framework with the lowest initial FBA, which rises from 63.09 to 69.59 on MIMIC-CXR. ReAct, whose errors produced the lessons, has the smallest gain on MIMIC-CXR and the second smallest on IU-Xray.

Table 4: **Lessons across agent frameworks and base model sizes.** Cells give the score with lessons; subscripts give the change from the same configuration without lessons. All rows use one store, built on MIMIC-CXR from errors of ReAct with Qwen3.5-4B; in size rows, one model writes the draft and applies the lessons. †Default configuration (Table 1).

	MIMIC-CXR				IU-Xray			
	RadGraph	CheXbert	GritLM	FBA	RadGraph	CheXbert	GritLM	FBA
Agent frameworks (Qwen3.5-4B)								
ReAct†	18.56$_{+1.80}$	53.79$_{+5.20}$	65.14$_{+1.32}$	67.69$_{+2.90}$	22.96$_{+2.39}$	41.50$_{+2.47}$	68.84$_{+1.18}$	59.63$_{+2.39}$
ReWOO	15.96$_{+0.22}$	51.85$_{+4.06}$	64.42$_{+0.91}$	69.43$_{+4.82}$	18.70$_{+0.07}$	39.86$_{+2.54}$	67.26$_{-0.01}$	59.13$_{+1.69}$
Plan-and-Solve	16.59$_{+2.42}$	52.06$_{+9.01}$	65.09$_{+1.38}$	69.59$_{+6.50}$	19.64$_{+3.74}$	40.48$_{+5.25}$	68.62$_{+0.77}$	59.26$_{+3.78}$
Swin-Qwen3	16.66$_{+1.09}$	52.32$_{+5.06}$	64.99$_{+1.11}$	69.40$_{+4.11}$	21.40$_{+3.40}$	40.80$_{+4.19}$	68.34$_{+0.18}$	59.62$_{+3.42}$
Base model sizes (ReAct, Qwen3.5)								
4B†	18.56$_{+1.80}$	53.79$_{+5.20}$	65.14$_{+1.32}$	67.69$_{+2.90}$	22.96$_{+2.39}$	41.50$_{+2.47}$	68.84$_{+1.18}$	59.63$_{+2.39}$
9B	17.75$_{+1.10}$	49.43$_{+0.88}$	63.91$_{+0.84}$	65.07$_{+0.35}$	21.22$_{+2.07}$	38.08$_{+0.48}$	68.52$_{+0.23}$	58.78$_{+1.90}$
27B	19.03$_{+2.56}$	54.54$_{+5.08}$	65.92$_{+1.51}$	69.67$_{+3.81}$	23.90$_{+2.03}$	42.95$_{+4.25}$	69.10$_{+0.18}$	59.69$_{+2.34}$
122B	19.39$_{+2.62}$	55.10$_{+1.99}$	66.30$_{+1.72}$	71.78$_{+3.21}$	24.13$_{+2.47}$	43.96$_{+4.52}$	69.70$_{+1.30}$	60.60$_{+2.04}$

Model sizes. Lessons written with Qwen3.5-4B improve agents built on base models of all four sizes, from 4B to 122B parameters (Table 4). All four clinical and semantic metrics rise in all eight comparisons across sizes and datasets, with FBA gains from 0.35 to 3.81 points. Lessons remain effective for models much larger than the one that wrote them: at 122B, they raise FBA by 3.21 points on MIMIC-CXR and 2.04 points on IU-Xray, giving the highest FBA among all sizes on both datasets (71.78 and 60.60). Lessons can also help a smaller model perform as well as a larger one. A 4B base model with lessons exceeds a 27B base model without them on both datasets, with FBA of 67.69 versus 65.86 on MIMIC-CXR and 59.63 versus 57.35 on IU-Xray.

4.7 Tracing report changes to lessons

Appendix D traces three test cases from the report change to the lesson, its images, and its writing cases, and checks each change by rerunning the policy without the lesson. In particular, for a correction case, the added finding disappears when the lesson is removed. For a case with an error introduced, the lesson was applied outside its stated condition, and its outcomes on the test set show 8 introduced errors and no corrections. So, the lesson can be identified and revised.

5 Conclusion and Limitations

We presented a vision-language memory framework that consolidates an agent's reporting errors into lessons, each pairing an explanation of a recurring error with both textual and image evidence behind it and links to its source episodes. On chest X-ray report generation, an MLLM agent by adopting 76 lessons consolidated from the episodes improved finding-level balanced accuracy more than retrieval over all 8,192 source episodes and more than two existing memory methods. The same set of lessons transferred without modification to IU-Xray, to three other agent frameworks, and to base models from 4B to 122B parameters with their performance enhanced, and lost part of their benefit when their images were removed or replaced. Training on repair and control cases increased corrections and reduced introduced errors with the same set of lessons. In our experiments, how experience was organized, and not only how much of it was stored, affected whether an agent could reuse it across data, agents, and models.

Lesson construction needs reference reports and Chest ImaGenome region boxes. The store is fixed during evaluation: we leave memory management, such as merging or retiring lessons, to future work, so a lesson that keeps introducing errors, as in Case 2, stays in the store. FBA depends on a language-model labeler (94.7% agreement with CAD-Chest on reference reports), and no radiologist has reviewed the generated reports, so the clinical value of the corrections is not established. Lessons come from one agent's errors on one hospital's data and were tested on one other chest X-ray dataset, using frontal images only; the 9B FBA gain on MIMIC-CXR is not distinguishable from zero. At inference, only the first followed lesson changes a report, so each case receives at most one lesson's correction.

AI USE STATEMENT

In this work, we used generative AI tools to help interpret our experimental results. This includes discussing how the results related to the paper's framing. We used AI coding assistants to help refactor and organize our code, including the experiment scripts, and to help prepare LaTeX tables. We have not used generative AI tools to generate synthetic data, formulate mathematical claims or proofs, or propose the research hypotheses; translation and qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the paper for structure and readability, to search for and summarize related work, and to format references. All numbers reported in the paper come from our own experiments. We reviewed all AI-assisted text, checked every cited paper against its source. The language models that are part of our method (the Qwen3.5 agent, the GLM-4.6V-Flash evaluation labeler) are described in Sections 3 and 4 and are separate from this disclosure. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses de-identified chest X-ray datasets. MIMIC-CXR, CAD-Chest, and Chest ImaGenome were accessed through PhysioNet under its credentialed data use agreement, and IU-Xray is publicly available. No new data were collected from human subjects and no additional institutional review was required. All models that processed MIMIC-CXR data ran on our own servers. The X-rays shown in Figures 1 and 2 come from these de-identified datasets. Our system is a research prototype that has not been validated for clinical use. Its lessons are built from a single hospital's data and can still introduce errors (Section 4.5), so generated reports should not be used without review by a radiologist. Because each lesson records the cases it was written from, its effect on a report can be traced and corrected; we view this as a support for future clinical evaluation, not as evidence of safety.

REPRODUCIBILITY STATEMENT

Section 3 describes lesson construction, retrieval, and training, and Section 4.1 gives the datasets, splits, models, and metrics. Appendix A gives the grouping criteria and the lesson-writer prompt, Appendix B describes how each agent framework receives lessons and lists data preparation and training hyperparameters. All base models and compared systems except Swin-Qwen3, which we reimplemented (Appendix B), are publicly available, and all datasets are available through PhysioNet or Open-i. We will release code, prompts, and lesson text upon publication. Because lesson images come from MIMIC-CXR, we will release them as study and image identifiers, so that credentialed users can reconstruct the full lessons.

REFERENCES

Hamida Abdaoui, Sabri Barbaria, Ahmed Al Kuwaiti, Noureddine Raouafi, Ridha Ben Salah, and Hanene Boussi Rahmouni. Swin-Qwen3: a three-stage vision–language framework for automated radiology report generation with multi-agent verification. *Frontiers in Radiology*, 6, 2026. ISSN 2673-8740. doi: 10.3389/fradi.2026.1875905. URL https://www.frontiersin.org/journals/radiology/articles/10.3389/fradi.2026.1875905.

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, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. MAIRA-2: Grounded radiology report generation, 2024. URL https://arxiv.org/abs/2406.04449v2.

Weihao Bo, Shan Zhang, Yanpeng Sun, Jingjing Wu, Qunyi Xie, Xiao Tan, Kunbin Chen, Wei He, Xiaofan Li, Na Zhao, Jingdong Wang, and Zechao Li. ViLoMem: Agentic learner with grow-and-refine multimodal semantic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5476–5486, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Bo_ViLoMem_Agentic_Learner_with_Grow-and-Refine_Multimodal_Semantic_Memory_CVPR_2026_paper.html.

Yurui Chang, Yiran Wu, Qingyun Wu, and Lu Lin. MemCollab: Cross-agent memory collaboration via contrastive trajectory distillation, 2026. URL https://arxiv.org/abs/2603.23234.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7346–7370, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.418. URL https://aclanthology.org/2024.emnlp-main.418/.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, Maya Varma, Andrew Johnston, Louis Blankemeier, Dave Van Veen, Tan Bui, Steven Truong, and Curtis Langlotz. RadGraph-XL: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12902–12915, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.765. URL https://aclanthology.org/2024.findings-acl.765/.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza M. Rodriguez, Sameer Kiran 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 : JAMIA*, 23 2:304–10, 2015.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 209–219. PMLR, 2021. URL https://proceedings.mlr.press/v158/endo21a.html.

Qianhan Feng, Zhongzhen Huang, Yakun Zhu, Yannian Gu, Winnie Chiu Wing Chu, Xiaofan Zhang, and Qi Dou. MedExpMem: Adapting experience memory for differential diagnosis, 2026. URL https://arxiv.org/abs/2605.22872.

Minghao Guo, Qingyue Jiao, Zeru Shi, Yihao Quan, Boxuan Zhang, Danrui Li, Liwei Che, Wujiang Xu, Shilong Liu, Zirui Liu, Mubbasir Kapadia, Vladimir Pavlovic, Jiang Liu, Mengdi Wang, Yiyu Shi, Dimitris N. Metaxas, and Ruixiang Tang. MemEye: A visual-centric evaluation framework for multimodal agent memory, 2026. URL https://arxiv.org/abs/2605.15128.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. Nguyen Duong, T. Bui, P. Chambon, M. Lungren, A. Ng, C. Langlotz, and P. Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports (version 1.0.0), 2021. URL https://doi.org/10.13026/hm87-5p47.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Subathra Adithan, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pp. 978–990. PMLR, 2024. URL https://proceedings.mlr.press/v227/jeong24a.html.

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, 2019. doi: 10.1038/s41597-019-0322-0. URL https://www.nature.com/articles/s41597-019-0322-0.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=BC4lIvfSzv.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652. Curran Associates, Inc., 2023. doi: 10.52202/075280-0377.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117/.

Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 643–655, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.28. URL https://aclanthology.org/2025.naacl-long.28/.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li,

Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2609–2634, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.147. URL https://aclanthology.org/2023.acl-long.147/.

Pan Wang, Yihao Hu, Xiujin Liu, Jingchu Yang, Hang Wang, and Zhihao Wen. AtlasVA: Self-evolving visual skill memory for teacher-free VLM agents, 2026. URL https://arxiv.org/abs/2605.17933v1.

Ziao Wang, Sixing Yan, Kejing Yin, Xiaofeng Zhang, and William K. Cheung. CURV: Coherent uncertainty-aware reasoning in vision-language models for x-ray report generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=3BTqvtwZYY.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63897–63911. PMLR, 2025b. URL https://proceedings.mlr.press/v267/wang25bx.html.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo Anthony Celi, Tanveer Syeda-Mahmood, and Mehdi Moradi. Chest ImaGenome dataset, 2021. URL https://physionet.org/content/chest-imagenome/1.0.0/. Version 1.0.0.

Peng Xia, Jianwen Chen, Hanyang Wang, Jiaqi Liu, Kaide Zeng, Yu Wang, Siwei Han, Yiyang Zhou, Xujiang Zhao, Haifeng Chen, Zeyu Zheng, Cihang Xie, and Huaxiu Yao. SkillRL: Evolving agents via recursive skill-augmented reinforcement learning, 2026. URL https://arxiv.org/abs/2602.08234v1.

Binfeng Xu, Zhiyuan Peng, Bowen Lei, Subhabrata Mukherjee, Yuchen Liu, and Dongkuan Xu. ReWOO: Decoupling reasoning from observations for efficient augmented language models, 2023. URL https://arxiv.org/abs/2305.18323.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12805–12825, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.583. URL https://aclanthology.org/2026.acl-long.583/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, Weizhi Zhang, Haodong Yue, and Wenya Wang. MemSkill: Learning and evolving memory skills for self-evolving agents, 2026a. URL https://arxiv.org/abs/2602.02474v2.

Mengliang Zhang, Xinyue Hu, Lin Gu, Tatsuya Harada, Kazuma Kobayashi, Ronald Summers, and Yingying Zhu. CAD-Chest: Comprehensive annotation of diseases based on MIMIC-CXR radiology report, 2023. URL https://physionet.org/content/cad-chest/1.0/. Version 1.0.

Yabin Zhang, Chong Wang, Yunhe Gao, Jiaming Liu, Maya Varma, Justin Xu, Sophie Ostmeier, Jin Long, Sergios Gatidis, Seena Dehkharghani, Arne Michalson, Eun Kyoung Hong, Christian Bluethgen, Haiwei Henry Guo, Alexander Victor Ortiz, Stephan Altmayer, Sandhya Bodapati, Joseph David Janizek, Ken Chang, Jean-Benoit Delbrouck, Akshay S. Chaudhari, and Curtis P. Langlotz. A reasoning-enabled vision-language foundation model for chest x-ray interpretation, 2026b. URL https://arxiv.org/abs/2604.00493.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936. URL https://ojs.aaai.org/index.php/AAAI/article/view/29936.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29733–29735, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, and Jun Wang. Memento: Fine-tuning LLM agents without fine-tuning LLMs, 2025. URL https://arxiv.org/abs/2508.16153.

A LESSON CONSTRUCTION DETAILS

This section describes how the lesson store used in the main experiments (76 lessons) was built. All steps use Qwen3.5-4B with greedy decoding.

Errors and groups. Lessons are built from the agent's no-memory drafts on 8,192 MIMIC-CXR studies from patients outside all evaluation sets. Reference-based feedback grades each finding in the CAD-Chest vocabulary. A finding that the reference states as present but the draft omits or only partly denies gives an *add* error; a finding that the reference states is absent but the draft asserts or hedges gives a *remove* error. Errors are grouped by finding and direction, with the number of groups per pair capped so that no single error type dominates. Each group has 16 cases from 16 distinct patients, and no patient appears in more than one group. The first 8 cases are the writing cases B_γ and the last 8 the check cases H_γ. Groups in which a writing case has no frontal image with Chest ImaGenome boxes are dropped, because the writer needs such an image to ground a lesson.

Writer. The writer runs in two stages. *Grounding* is one call per writing case: given the target finding and direction, the reference report, the draft, the feedback, and the frontal image with its list of Chest ImaGenome regions, the writer selects one region from the list as visible support, or declines. It cannot propose its own coordinates. *Synthesis* is one call per group: given the grounding notes and the grounded images with their regions drawn, the writer selects one to three images from distinct patients and writes the lesson's condition, cue and exception. The decision field follows from the direction. A response that breaks the output format is retried once; the model's text is never edited. The system prompts follow.

Grounding system prompt.

```
Review one completed chest-radiograph writing case. Its reference report and
feedback are teaching evidence about this earlier case, never about a future patient.
Select a visible anatomical region that supports a reusable cue for the named target
correction. Use only the supplied image_id and region_id; do not generate coordinates.
Read the pixels as well as the wording. Preserve uncertainty, temporal qualifications,
and alternative explanations. Explain when the cue applies and when it does not.
Return JSON only with exactly these keys:
{"usable":true,"image_id":"menu image ID","region_id":"menu region ID",
"summary":"visible applicability pattern","cue":"visible discriminator",
"exception":"when this cue should not be used"}.
If there is no credible support, use usable=false, empty image_id and region_id,
and explain the limitation in summary; cue and exception may be empty.
Keep each text field at most 400 characters.
The target direction is an instruction, not a diagnosis to repeat:
- add: find observable support for correcting omission or understatement of the named
  target.
- remove: find observable grounds for withdrawing or qualifying an unsupported assertion of
  the named target.
A coexisting finding, unrelated anatomical region, or absence of a mention in the draft
does not establish this correction. If the image cannot support the named correction,
return usable=false. In particular, evidence of another organ being normal cannot
establish that the target abnormality is absent.
The reference explains the historical teaching error. Only summary may describe that
historical error. Write cue and exception as visual tests for another patient's images;
they must not depend on that future patient's reference report or diagnosis label.
```

Synthesis system prompt.

```
Write one reusable visual lesson for the named target correction.
BUILD references and feedback explain earlier patients' errors. The future reporter
has only its radiographs, indication and draft. Write a condition, visible cue and
exception it can check from those inputs. Draft wording is not image evidence.
For add, teach observable support for restoring an omitted or understated target.
For remove, teach grounds for withdrawing or qualifying an unsupported target assertion.
Preserve uncertainty, temporal qualifications and alternative explanations; preserve
non-target findings. Discard notes unsupported by the images or original case evidence.
Each candidate image is immediately preceded by its support ID and case note.
Select one to three informative, mutually compatible supports from distinct patients.
Return JSON only, with supports FIRST and all five fields:
{"supports":["support ID"],"usable":true,"condition":"when to apply",
"cue":"visible discriminator","exception":"when not to apply"}.
Use short sentences, at most 400 characters per text field. Never refer to reference
reports, labels, grading or source patient IDs in the three lesson text fields.
If no transferable visual rule is supported, return supports=[], usable=false,
a brief reason in condition, cue="", exception="". Empty exception is otherwise
```

```
allowed only when no defensible exception is available. Do not invent one.
Separate learning from application. Every following case note describes a PAST
BUILD patient. Its reference, feedback and grounding summary help you learn;
they are not facts available about a future patient.
In condition, describe the future image pattern and relevant draft stance.
In cue, describe the visible evidence that distinguishes the target.
In exception, describe an observable alternative or limitation.
Structural example only, not a finding to copy: condition='When the draft omits
the target and the current image shows [supported pattern]'; cue='[visible
discriminator]'; exception='Do not apply when [observable alternative or limitation]'.
Replace brackets with evidence supported by these images. If that cannot be done,
abstain. Do not turn agreement with a past reference into a future application test.
```

Lesson images. A lesson image is the full frontal radiograph of a writing case, resized to 518×518, with the selected Chest ImaGenome box drawn on it. We keep lesson images for 14 regions: the upper, mid and lower lung zones, the costophrenic angles, the hilar structures and the lungs on each side, plus the cardiac silhouette and the upper mediastinum. The writer selects one to three of these images, at most one per region.

B Experimental setup details

B.1 Data

MIMIC-CXR. We start from all MIMIC-CXR studies whose findings and impression sections each have more than 10 words and that have at least one image (76,066 studies, 31,201 patients). We split *patients*, not studies, because a patient has more than two studies on average. Whole patients are then assigned, alternately, to a validation split and a test split until each holds 1,000 studies. All remaining patients form the construction stream.

Split	Studies	Patients	Used for
Construction stream	74,066	30,422	lesson construction, policy training
Validation	1,000	385	development and checkpoint selection
Test	1,000	394	reported results

No patient appears in two splits; we checked this directly on the index. Lessons are built from the first 8,192 studies of the stream. Policy training uses stream patients only, and explicitly excludes every patient of the validation and test splits. The primary image is the PA view, else AP.

IU-Xray. We use 1,000 IU-Xray studies that have a frontal image and a non-empty findings section. Each study contributes one frontal image; lateral views are not used. The reference text is the findings and impression joined.

B.2 Compared systems and agent frameworks

All compared systems run on our own servers from their public checkpoints, with the same test studies, inputs, and evaluation pipeline as our agent, and all methods receive images at 518×518 resolution. For a given agent, all memory conditions run the same model. Episodic memory and lessons also share prompts, settings, and the acceptance rule, and differ only in what is retrieved; ExpeL and ViLoMem are described below. Episodic memory retrieves individual source episodes from the same 8,192 construction studies used to build lessons. Swin-Qwen3 has no public code; we reimplement its three prompted agents (drafting, clinical verification, refinement) following the paper's descriptions, using Qwen3.5-4B for all three in place of its trained Swin/Q-Former visual interface, because Qwen3.5-4B reads images directly, and greedy decoding rather than beam search. For the model-size experiments, the lesson store written by Qwen3.5-4B is kept unchanged. The agent of each size writes its own draft and applies the lessons with the same model, using the base policy without training.

ExpeL and ViLoMem. Both are ported from their public code and built from the same 608 writing cases as the lesson store, with the frozen Qwen3.5-4B as writer and no access to evaluation labels. ExpeL (Zhao et al., 2024) distills the cases into a list of global rules with the original ADD/EDIT/AGREE/REMOVE scoring and cap of 10; 6 rules remain. At inference all rules are

given with the reference reports of the two most similar construction cases. ViLoMem (Bo et al., 2026) writes one logical and one visual guideline per writing case with the original prompts (546 and 524 guidelines), retrieves the top 3 of each by text and image similarity, and places them in the original guideline template; its merge and refine step and attention heatmaps are not connected in the released code and are not used. Each method receives the draft in one revision call, and the revised report is the final report. Qwen3.5-4B replaces the GPT and Qwen-VL APIs; our retrieval embedding replaces all-mpnet-base-v2 for ExpeL, and bge-small-en-v1.5 and RAD-DINO replace Qwen3-Embedding-8B and the DashScope embedding for ViLoMem.

B.3 CONFIDENCE INTERVALS.

We estimate uncertainty with a case-level percentile bootstrap. We draw n cases with replacement from the n test cases, 5,000 times, with a fixed seed, and use the same 5,000 draws for every method and metric on a dataset. The 95% confidence interval of a score is given by the 2.5th and 97.5th percentiles of its 5,000 bootstrap values; tables report its half-width, since the two sides differ by at most 0.15 points. To compare two methods, we compute the difference between their scores on each draw and take the 2.5th and 97.5th percentiles of the 5,000 differences. Because both methods are scored on the same cases in each draw, this interval accounts for the pairing between methods.

B.4 REPORTING POLICY

Prompt. The agent first writes the draft; for the trained policy, the draft is also written with the trained model. Each retrieved lesson is then used in a separate call, which holds the current-patient images and the agent's pre-report observations exactly as the agent saw them, the draft, and one lesson. The lesson images are full frontal radiographs of writing cases, with the lesson's Chest ImaGenome box and region name drawn on them. Lessons are tried in retrieval order ($K = 3$). The first revision that moves the lesson's target finding in the lesson's direction becomes the final report; if none does, the draft is the final report. Decoding is greedy. The system prompt and the user message are:

System.

```
You are a radiologist writing the report for one chest radiograph.
Use only the current patient's indication, radiographs, and accumulated observations in the
  user
message. Cover the lungs, pleura, heart and mediastinum, and any devices supported by the
  images.
There is no prior radiograph, so do not describe change, resolution, persistence, or
  comparison.
Never infer current-patient evidence from historical text or images. Return only a complete
  report
with FINDINGS and IMPRESSION sections and no Thought, Action, or Observation.
```

User message (bracketed items are images or inserted fields).

```
[current-patient images and the agent's pre-report observations, unchanged]

Pre-memory draft from this patient:
{draft}
End of pre-memory draft.

Historical lesson from earlier patients - not evidence about this patient.

Apply this lesson only if its condition and visual cue are supported by the
current patient's images and accumulated observations. Ignore it when it is
inapplicable or an exception is present. Historical images are analogies only.

Condition: {condition}
Reporting decision: {decision}
Visual cue: {cue}
Exceptions: {exception}

[one to three lesson images]
End of historical lesson.

Write a complete replacement report. Return only 'FINDINGS: ...' followed by 'IMPRESSION:
  ...'.
```

B.5 RETRIEVAL

The embedding $h(\xi)$ is computed by the base Qwen3.5-4B itself. Its input has images and a short JSON text. The images are the primary radiograph, and the last crop or lateral view the agent opened (or the primary radiograph again if it opened none). The text holds the indication, the pre-report observations, and the draft. $h(\xi)$ is the last-layer hidden state of the final token, L_2-normalized. We then subtract the mean embedding of the first 256 construction-stream studies and normalize again. A lesson's key k_γ is the normalized mean of h over its 8 writing cases. The top three lessons by cosine similarity are retrieved.

B.6 TRAINING

Prompts. Training prompts pair a lesson from the lesson store (Appendix A) with a stream case. A *repair* prompt is a case where the draft is wrong about the lesson's target in the way the lesson corrects: the reference states the target and the draft misses it (add), or the reference rules it out and the draft asserts it (remove). A *control* prompt is a case where the draft already agrees with the reference on the target and following the lesson would make it wrong. Repair and control prompts are balanced; the 300 updates used 3,600 prompts (1,797 repair, 1,803 control).

Reward. The reward is computed on the final report. It is $+1$ for each graded finding that becomes a true positive or stops being a false positive, and -1 for each change in the other direction. A malformed response receives -1.

Optimization. GRPO in ms-swift (Zhao et al., 2025), starting from the frozen Qwen3.5-4B. Advantages are normalized by the group mean and standard deviation (scale_rewards=group). Importance ratios are sequence-level (importance_sampling_level=sequence), following GSPO. Loss aggregation follows Dr. GRPO (loss_type=dr_grpo) (Liu et al., 2025): token-level losses are summed and divided by the number of samples times the maximum completion length, so no per-response length normalization is applied. Note that we adopt only the loss aggregation of Dr. GRPO and only the sequence-level ratio of GSPO; the standard-deviation normalization of GRPO is kept.

B.7 METRICS

- **FBA.** We label each report over the 27-finding CAD-Chest vocabulary with GLM-4.6V-Flash (V Team et al., 2025). The model reads the report text and states, for each finding, whether the report asserts, hedges, denies, or does not mention it. Asserted and hedged findings count as present; denied, partly denied, and unmentioned findings count as absent. Reference labels come from CAD-Chest on MIMIC-CXR, CAD-Chest does not release its extraction code, so its rules cannot be applied to generated reports, and we label generated reports with GLM-4.6V-Flash instead. On the 1,000 MIMIC-CXR test reference reports, the labeler's present/absent decisions agree with CAD-Chest on 94.7% of labeled study-finding pairs ($\kappa = 0.89$); on IU-Xray, we label the reference reports in the same way. The labeler runs locally with greedy decoding. The calculation of FBA is as follows: Let i index the test studies and f the 27 findings. For study i, $\hat{y}_i$ is the generated report, $p_f(\hat{y}_i) \in \{0, 1\}$ is the extracted label for finding f as defined above, and $t_{f,i} \in \{0, 1\}$ is the reference label; $\mathbf{1}\{\cdot\}$ is the indicator function. Counts are pooled over all study-finding pairs:

$$\mathrm{TP} = \sum_{i,f} \mathbf{1}\{p_f(\hat{y}_i) = 1,\ t_{f,i} = 1\}, \qquad \mathrm{FN} = \sum_{i,f} \mathbf{1}\{p_f(\hat{y}_i) = 0,\ t_{f,i} = 1\},$$
$$\mathrm{FP} = \sum_{i,f} \mathbf{1}\{p_f(\hat{y}_i) = 1,\ t_{f,i} = 0\}, \qquad \mathrm{TN} = \sum_{i,f} \mathbf{1}\{p_f(\hat{y}_i) = 0,\ t_{f,i} = 0\}.$$

$$\mathrm{Sens} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{Spec} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \qquad \mathrm{FBA} = \tfrac{1}{2}\left(\mathrm{Sens} + \mathrm{Spec}\right).$$

- **RadGraph.** F1RadGraph from the `radgraph` package (v0.1.18) with the RadGraph-XL (Delbrouck et al., 2024) model, *complete* reward level: an entity counts only if its span, its label and all its relations match. Scores are averaged over studies.

- **CheXbert.** F1CheXbert (`f1chexbert` package). CheXbert labels 14 CheXpert observations in each whole report, generated and reference. Positive and uncertain count as positive. We report micro-averaged F1 over all 14 labels, computed once over the whole test set.

- **GritLM.** GritLM-7B embeds each whole report with the instruction "encode the medical text:". We report the cosine similarity between generated and reference report, averaged over studies.

- **BLEU-4, ROUGE-L, METEOR.** Per report, then averaged. BLEU-4 is NLTK `sentence_bleu` with smoothing method 1, ROUGE-L is the F-score of the `rouge` package, and METEOR is NLTK `meteor_score`.

C PAIRED COMPARISONS

Tables 5 and 6 report paired differences for the comparisons that support the main conclusions of Section 4. Each entry is the score of the first configuration minus the score of the second, in points on the 0–100 scale, with its 95% interval below it. Intervals come from the case-level bootstrap of Section 4.1: in each of the 5,000 draws, both configurations are scored on the same resampled test cases, and the interval is given by the 2.5th and 97.5th percentiles of the 5,000 differences. All entries for a dataset use the same draws. Entries whose interval contains zero are shown in grey.

Lessons raise all four clinical and semantic metrics over the agent without memory on both datasets, with every interval excluding zero. Removing the lesson images, removing the lesson text, or replacing the lesson images with randomly sampled chest X-rays lowers all four metrics on both datasets, with every interval excluding zero. In the three other agent frameworks, lessons raise FBA and CheXbert on both datasets with intervals that exclude zero. Across base model sizes, the FBA interval excludes zero in every case except the 9B model on MIMIC-CXR. For policy training (Table 6), the trained policy exceeds the base policy on all four metrics, and removing the lesson images from the trained policy lowers FBA by 1.41 points ($[-2.07, -0.72]$).

Table 5: **Paired differences for the main comparisons.** Each entry is the first configuration minus the second, in points; the 95% case-level paired bootstrap interval is given below it (Section 4.1). Grey entries have an interval that contains zero. Unless stated otherwise, rows use the ReAct agent on Qwen3.5-4B with the base policy. Framework and size rows give the change from adding lessons to that configuration; the ReAct and 4B rows equal *Lessons − no memory*.

	MIMIC-CXR				IU-Xray			
Comparison	RadGraph	CheXbert	GritLM	FBA	RadGraph	CheXbert	GritLM	FBA
Lessons (RQ1, RQ2)								
Lessons − no memory	+1.80 [1.38, 2.25]	+5.20 [4.05, 6.36]	+1.32 [1.09, 1.55]	+2.90 [2.04, 3.76]	+2.39 [1.96, 2.83]	+2.47 [1.06, 3.90]	+1.18 [1.01, 1.37]	+2.39 [1.63, 3.15]
Lesson content (RQ3; condition − lessons)								
w/o visual evidence	−1.47 [−1.87, −1.09]	−2.43 [−3.61, −1.31]	−0.84 [−1.06, −0.61]	−1.48 [−2.40, −0.55]	−0.90 [−1.31, −0.49]	−1.61 [−3.00, −0.19]	−0.62 [−0.79, −0.45]	−1.61 [−2.40, −0.82]
w/o lesson text	−2.49 [−2.90, −2.08]	−3.89 [−5.08, −2.71]	−1.32 [−1.56, −1.09]	−2.19 [−3.12, −1.28]	−1.04 [−1.49, −0.62]	−2.91 [−4.34, −1.44]	−0.75 [−0.91, −0.57]	−2.79 [−3.61, −1.95]
random images	−1.67 [−2.08, −1.27]	−3.09 [−4.24, −1.93]	−1.02 [−1.25, −0.79]	−2.20 [−3.13, −1.25]	−1.16 [−1.55, −0.78]	−2.80 [−4.23, −1.37]	−0.74 [−0.92, −0.56]	−1.96 [−2.80, −1.10]
Agent frameworks, Qwen3.5-4B (RQ5)								
ReWOO	+0.22 [−0.32, 0.78]	+4.06 [2.50, 5.69]	+0.91 [0.60, 1.23]	+4.82 [3.56, 6.12]	+0.07 [−0.66, 0.83]	+2.54 [0.60, 4.45]	−0.01 [−0.32, 0.29]	+1.69 [0.52, 2.90]
Plan-and-Solve	+2.42 [1.80, 3.02]	+9.01 [7.32, 10.75]	+1.38 [1.07, 1.68]	+6.50 [5.19, 7.82]	+3.74 [3.04, 4.43]	+5.25 [3.37, 7.15]	+0.77 [0.54, 0.99]	+3.78 [2.57, 5.03]
Swin-Qwen3	+1.09 [0.51, 1.64]	+5.06 [3.53, 6.62]	+1.11 [0.81, 1.41]	+4.11 [2.86, 5.33]	+3.40 [2.71, 4.09]	+4.19 [2.54, 5.88]	+0.18 [−0.05, 0.41]	+3.42 [2.26, 4.55]
Base model sizes, ReAct (RQ5)								
Qwen3.5-9B	+1.10 [0.68, 1.52]	+0.88 [−0.34, 2.08]	+0.84 [0.61, 1.07]	+0.35 [−0.56, 1.26]	+2.07 [1.79, 2.36]	+0.48 [−0.69, 1.67]	+0.23 [0.06, 0.40]	+1.90 [1.26, 2.52]
Qwen3.5-27B	+2.56 [2.13, 2.97]	+5.08 [3.82, 6.34]	+1.51 [1.30, 1.74]	+3.81 [2.85, 4.78]	+2.03 [1.53, 2.53]	+4.25 [2.95, 5.52]	+0.18 [−0.03, 0.39]	+2.34 [1.55, 3.16]
Qwen3.5-122B	+2.62 [2.19, 3.05]	+1.99 [0.96, 3.05]	+1.72 [1.51, 1.91]	+3.21 [2.35, 4.07]	+2.47 [1.78, 3.16]	+4.52 [3.27, 5.82]	+1.30 [1.01, 1.59]	+2.04 [1.37, 2.74]

Table 6: **Paired differences for policy training on MIMIC-CXR.** Same format as Table 5. Both rows use the same lesson store and retrieval (Table 3).

Comparison	RadGraph	CheXbert	GritLM	FBA
Trained − base (both with lessons)	$+0.24$ [0.03, 0.46]	$+3.05$ [2.20, 3.89]	$+1.31$ [1.18, 1.44]	$+2.28$ [1.60, 2.97]
Trained: w/o visual − lessons	-0.16 [−0.36, 0.03]	-2.10 [−2.84, −1.40]	-1.19 [−1.32, −1.06]	-1.41 [−2.07, −0.72]

D CASE STUDY: TRACING REPORT CHANGES TO LESSONS

This appendix shows how a change in a generated report can be traced to the lesson that caused it and to the cases the lesson was written from (Figure 3). All cases come from the trained reporting policy on the MIMIC-CXR test set, and every value is read from stored outputs. Three properties of the framework make the trace possible. First, each lesson is tried in a separate call with the draft, so a change in the final report comes from one lesson. Second, a counterfactual run with the same policy and draft, but with the lesson replaced by a content-free block, shows whether the change disappears without the lesson. Third, each lesson links to its writing cases and lesson images, and to its outcomes on the cases where it was tried.

Case 1: a correction. Three lessons were retrieved. The first left the draft unchanged, and the second edited the report without changing its target finding in the lesson's direction, so neither was accepted. The third lesson added edema, the only change to the report, and the counterfactual returned the draft unchanged, so the added finding comes from this lesson. The trace shows what the lesson states, the image evidence it was written from, and how it behaved elsewhere: on the test set it was followed on 11 cases, correcting 5 and introducing 1 error. The reference also reports small pleural effusions, which the final report still misses; this case concerns edema only.

Case 2: an introduced error. Here the trace explains an error. The lesson applies when a draft omits an effusion or describes the costophrenic angle as sharp. This draft explicitly denies effusion, so the trained policy applied the lesson outside its condition; the untrained policy declined the same lesson on the same case. The lesson's record shows that this is not an isolated failure: on the test set it was followed on 11 cases, introducing 8 errors and correcting none. Its check record did not reveal the problem, because all 8 of its check cases have effusion present, so a false effusion could not occur there. The trace thus points to two improvements that we leave to future work: adding check cases in which the lesson's finding is absent, and removing lessons with no demonstrated correction.

Case 3: lessons correctly not applied. Three lessons were retrieved for a film whose draft and reference both describe clear lungs, and the policy returned the draft unchanged for each. For the effusion lesson, the condition requires opacity obscuring the right hemidiaphragm and costophrenic angle, which neither the draft nor the reference describes.

Case 1: correction (random draw). Study 55233514, AP. Target finding: edema. Reference: *"Mild congestive heart failure with cardiomegaly, mild pulmonary edema. . . "*



current image

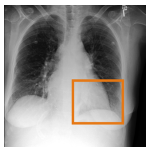
Draft (no lessons): *"IMPRESSION: Cardiomegaly with bilateral pulmonary opacities, most prominent in the left lower lobe. . . "* Edema is not mentioned.

Lessons tried, in retrieval order:
1. add lung opacity: draft unchanged; not followed.
2. add pleural effusion: report edited, but its target did not move in the lesson's direction; not followed.
3. add edema: edema changed from unmentioned to asserted; **followed**.

Final report: *"IMPRESSION: Cardiomegaly **with pulmonary edema** and bilateral pulmonary opacities. . . "* The only change in the report. Correct for edema.

Counterfactual: with the lesson replaced by a content-free block, the policy returns the draft unchanged.

Trace: lesson "add edema"



lesson image (writing case 58365028)

Condition: the draft omits edema and the image shows increased interstitial markings, peribronchial cuffing, or hazy lower-zone opacification.

Exception: markings due to chronic interstitial disease or fibrosis.

Source: written from 8 writing cases (50504286, 54839826, 58365028, . . .).

Test-set record: tried on 25 cases, followed on 11: 5 corrections, 1 introduced error.

Case 2: introduced error (chosen by hand). Study 56783999, PA. Target finding: pleural effusion. Reference: *"There are no pleural effusions."*



current image

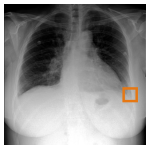
Draft (no lessons): *"There is no evidence of pleural effusion or pneumothorax."* Correct.

Lessons tried, in retrieval order:
1. add pleural effusion: effusion changed from denied to asserted; **followed**.

Final report: *"There is no evidence of pneumothorax. **There is blunting of the right costophrenic angle, consistent with a small right pleural effusion.**"* Wrong.

Counterfactual: with the lesson replaced by a content-free block, the policy keeps the correct draft sentence.

Trace: lesson "add pleural effusion"



lesson image (writing case 59653923)

Condition: the draft *omits* an effusion or states that the costophrenic angle is sharp. This draft explicitly denies effusion, so the lesson was applied outside its condition.

The untrained policy declined this lesson on this case.

Test-set record: tried on 28 cases, followed on 11: 8 introduced errors, 0 corrections.

Check record: all 8 check cases have effusion present, so the check could not reveal false positives.

Case 3: lessons retrieved but not applied (chosen by hand). Study 57556209, PA. Target finding: pleural effusion. Reference: *"There is no evidence of pneumomediastinum or effusion."*



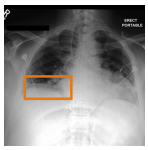
current image

Draft (no lessons): *"The lungs are clear without focal consolidation, pleural effusion, or pneumothorax."* Correct.

Lessons tried, in retrieval order:
1. add vascular congestion: draft unchanged; not followed.
2. add pleural effusion: draft unchanged; not followed.
3. add atelectasis: draft unchanged; not followed.

Final report: the draft, unchanged. Correct.

Trace: lesson "add pleural effusion"



lesson image (writing case 56537307)

Condition: the image shows opacity in the right lower lung zone obscuring the right hemidiaphragm and costophrenic angle. Both the draft and the reference describe clear lungs, so the condition does not hold.

Figure 3: **Tracing report changes to lessons.** For each case we show the lessons tried and whether each was followed, the change to the report, a counterfactual in which the lesson is replaced by a content-free block, and the trace from the applied lesson to its text, a lesson image, its writing cases, and its record on the test set. Lesson text is abbreviated; boxes on lesson images are Chest ImaGenome regions.