STRIVE: Multi-Agent Structured Temporal Reasoning with Integrated Verification for Longitudinal Radiology Report Generation

Junyeong Maeng[1], **Eunsong Kang**[2*], **Heung-Il Suk**[1*]

[1]Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
[2]Graduate School of Data Science, Kangwon National University, Chuncheon, Republic of Korea
mjy8086@korea.ac.kr, eskang@kangwon.ac.kr, hisuk@korea.ac.kr

Abstract

Longitudinal radiology report generation (LRRG) requires identifying both current findings and their changes relative to a prior study. Existing methods jointly model diagnosis, attribute estimation, temporal comparison, and language generation within implicit representations, which can cause task interference, obscure the evidence underlying each decision, and limit error traceability. They also model progression states as independent labels, ignoring their ordered structure and thus treating missed changes and direction reversals equally. We present STRIVE, Multi-Agent Structured Temporal Reasoning with Integrated Verification for LRRG, which decomposes clinical reasoning into specialized Diagnosis, Attribute, and Temporal Change Agents that produce explicit intermediate evidence. In particular, the Temporal Change Agent is further post-trained using Progression-Aware GRPO, a verifiable, shaped reward that assigns partial credit to direction-preserving errors while scoring direction reversals lowest. STRIVE performs verification at two stages: a deterministic Consistency Gate reconciles the agent outputs before report generation, and a Validation Agent checks whether the generated report is supported by the aggregated clinical evidence. On Longitudinal-MIMIC, STRIVE attains the best clinical efficacy among recent methods and more than doubles Longitudinal Change Concordance (LCC), a measure of temporal agreement with the reference report, over the strongest baseline.

Introduction

Radiology report generation (RRG) aims to automatically generate clinically meaningful reports from medical images. It is a challenging task because models must accurately identify fine-grained clinical information, including disease presence, severity, and location, and faithfully express it in natural language (Brady 2018). Recent advances in vision-language models and large language models have improved the linguistic fluency and clinical accuracy of generated reports, driving active research in RRG (Wang et al. 2023, 2024; Liu et al. 2024). More recently, this task has been extended to longitudinal radiology report generation (LRRG), which compares studies across multiple time points to capture disease progression beyond findings observed at a single time point (Wang, Du, and Yu 2024; Nicolson et al. 2024; Liu et al. 2026a) as illustrated in Figure 1(a).

Recent LRRG methods incorporate longitudinal information in different ways. Alignment-based approaches establish temporal correspondences between prior and current studies by aligning visual or vision-language features (Liu et al. 2026b; Gao et al. 2026), whereas fusion-based approaches integrate features across studies into a unified longitudinal representation for report generation (Wang, Du, and Yu 2024; Dong et al. 2026), as illustrated in Figure 1(b). Despite effectively incorporating comparative information, these approaches face two key challenges: the implicit coupling of clinical reasoning and report generation, and inadequate modeling of directional relationships among progression states.

The first limitation stems from implicit coupling of diagnosis, clinical attribute estimation, temporal reasoning, and report generation within a shared representation. These tasks impose different representational demands. Report generation favors a smooth semantic space, whereas clinical objectives require a discrete and constrained decision space. Their joint optimization can cause task interference, obscure fine-grained evidence, and distort inferred information during generation (Zhang et al. 2026a). Moreover, when these rea-

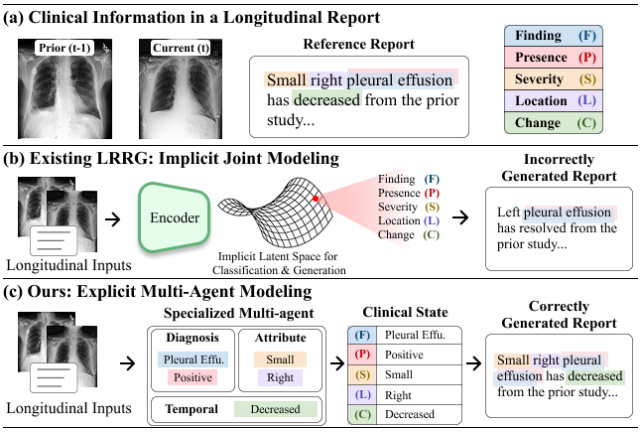

Figure 1: (a) A longitudinal report includes finding, presence, severity, location, and change. (b) Existing LRRG entangles them in an implicit latent space, generating clinically inconsistent reports. (c) STRIVE decomposes them across specialized agents, forms an explicit clinical state, and verifies the final report.

*Corresponding author.

soning processes are exposed only through the final report, it becomes difficult to determine whether an error originates from clinical reasoning or language generation. Explicit intermediate evidence is therefore essential for error localization, analysis, and clinical verification (Tanida et al. 2023).

The second limitation lies in inadequate modeling of relationships among progression states. Although recent LRRG methods use progression labels or change descriptions as supervision (Yun et al. 2025; Dong et al. 2026), they typically model these states as independent targets. This overlooks their directional structure: `new` and `resolved` represent opposite transitions, while `increased`, `stable`, and `decreased` reflect ordered changes in persistent findings. Consequently, conventional objectives do not adequately distinguish a missed progression, such as predicting `stable` instead of `increased`, from a direction reversal, such as predicting `decreased` instead of `increased`, despite their different clinical and directional severity.

Together, these limitations motivate an explicit multi-agent framework (Khan et al. 2026; Zhang et al. 2026b) that separates diagnosis, attribute estimation, and temporal reasoning across specialized agents, enabling task-specific optimization while making intermediate clinical evidence explicit. However, agent specialization may introduce cross-agent inconsistency, where agents produce conflicting findings, and evidence-to-report unfaithfulness, where valid intermediate evidence is omitted or distorted during report generation. Reliable multi-agent LRRG therefore requires integrated verification to keep the final report grounded in aggregated clinical evidence.

Alongside these modeling considerations, evaluation should also reflect the temporal correctness of generated reports. Existing diagnosis-based (Smit et al. 2020), lexical (Papineni et al. 2002; Lin 2004; Banerjee and Lavie 2005), model-based (Zhang et al. 2020; Ostmeier et al. 2024), and structure-based (Jain et al. 2021; Yu et al. 2023) metrics capture complementary aspects of report quality, including disease accuracy, linguistic similarity, semantic agreement, and structural consistency. However, progression-specific properties are assessed only indirectly. A more direct evaluation of disease-specific change coverage and directional agreement is therefore needed for LRRG.

To address these limitations, we propose STRIVE: Multi-Agent Structured Temporal Reasoning with Integrated Verification for Longitudinal Radiology Report Generation. STRIVE decomposes longitudinal interpretation into specialized clinical reasoning agents, explicitly models progression-state relationships, and integrates evidence reconciliation with report-level validation as illustrated in Figure 1(c). Our contributions are fourfold:

- We propose STRIVE, a multi-agent LRRG framework that decomposes diagnosis, clinical attribute estimation, and temporal progression modeling into specialized agents with explicit intermediate evidence.
- We introduce Progression-Aware GRPO, a reinforcement learning with verifiable rewards (RLVR) objective whose shaped, multi-level reward captures the directional and graded relationships among longitudinal progression states (*e.g.*, increased, decreased, stable).
- We develop a Consistency Gate to resolve logical conflicts among agent outputs and a Validation Agent to verify that the final report is supported by the committed clinical state.
- We evaluate STRIVE on Longitudinal-MIMIC and show that it outperforms existing methods in linguistic fluency, diagnostic performance, and progression-state agreement measured by Longitudinal Change Concordance (LCC).

Related Work

Longitudinal Radiology Report Generation

Conventional RRG primarily focuses on recognizing clinical findings, including disease presence, severity, and location, from a single study (Chen et al. 2020, 2021; Wang et al. 2023; Jin et al. 2024). However, methods based on a single study do not directly model cross-time disease changes, which require comparisons between prior and current studies. LRRG extends RRG by incorporating historical studies and modeling temporal evidence.

Existing LRRG methods differ mainly in how they relate the prior study to the current one. Fusion-based methods, such as PriorRG (Liu et al. 2026a), integrate prior and current visual features into a unified spatiotemporal representation. Alignment-based methods explicitly establish cross-time relationships between local visual regions. BiOTPrompt (Liu et al. 2026b) employs bidirectional optimal transport to identify asymmetric patch-level changes, whereas MARE (Gao et al. 2026) dynamically aligns lesion regions and performs analogical reasoning over visual and textual evolution relations. TIM (Dong et al. 2026) instead separates the spatial representation of each finding from the modeling of how it progresses between the two studies.

Despite their architectural diversity, existing LRRG methods encode disease recognition, clinical attribute estimation, and temporal comparison primarily within shared implicit representations. Such joint modeling causes interference among heterogeneous reasoning tasks, making errors difficult to trace and verify.

Multi-Agent Clinical Reasoning and Verification

Recent RRG studies have begun to adopt multi-agent formulations that structure image interpretation and report generation into specialized stages based on clinical workflows. For example, CogRad (Khan et al. 2026) assigns global triage, regional investigation, report writing, and output verification to Scout, Investigator, Writer, and Verifier agents, respectively. RadAgents (Zhang et al. 2026b) divides chest X-ray interpretation by anatomical region and integrates the per-region analyses with a Synthesizer agent. Both also verify their outputs, CogRad by re-examining sentences that lack sufficient visual support and RadAgents by resolving inconsistencies across agent and tool outputs. These formulations, however, decompose the interpretation of a single study, so no agent is responsible for deciding how a finding has changed relative to a prior one. STRIVE makes that decision an explicit agent output and verifies it against the diagnosed presence before the report is written.

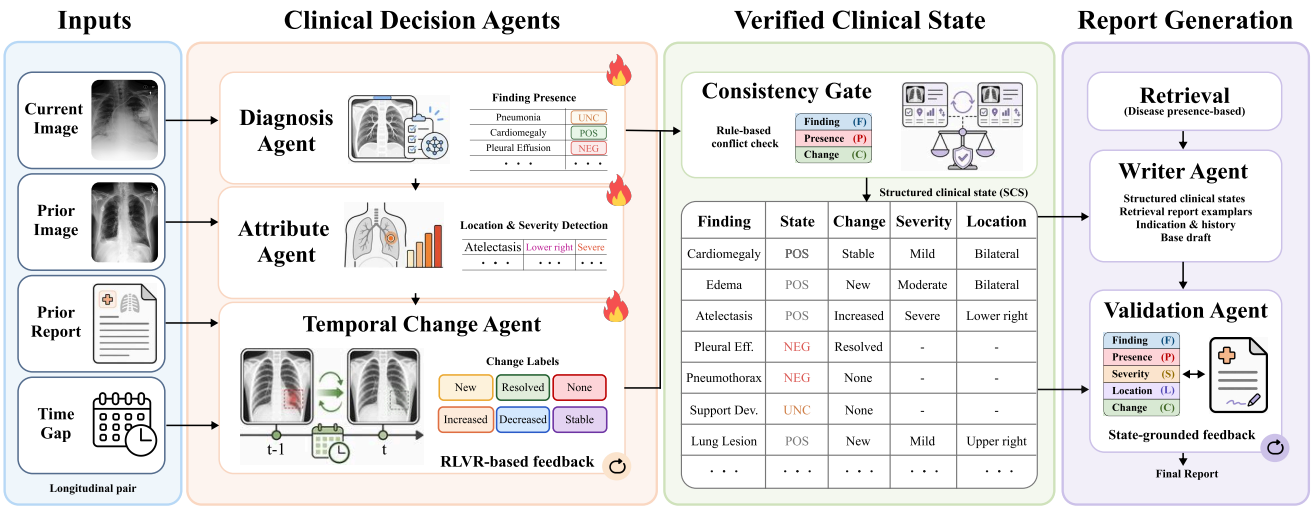

Figure 2: Overview of STRIVE. Three Clinical Decision Agents infer per-finding presence (Diagnosis), severity and location (Attribute), and change labels (Temporal Change) post-trained with GRPO. A rule-based Consistency Gate resolves diagnosis–change conflicts and aggregates the outputs into a structured clinical state. A Writer Agent then generates the report and a Validation Agent applies state-grounded edits to produce the final report.

Method

Framework Overview

Let $X_j = (I_j, v_j)$ denote the chest X-ray image I_j and its acquisition view v_j at time $j \in \{t-1, t\}$. The most recent prior study also provides its report r_{t-1} and the inter-study interval δ_t. We use c_t for the available clinical context, restricted to the study history and indication. Our framework, STRIVE, generates the current report $\hat{r}_t$ as

$$\hat{r}_t = \mathcal{G}(X_t, X_{t-1}, r_{t-1}, \delta_t, c_t). \tag{1}$$

We apply task decomposition to longitudinal report generation, factorizing it into three clinical decisions: presence in the current study, attributes of a present finding, and change relative to the prior study. Each decision is handled by a role-specialized agent. The *Diagnosis Agent* determines whether each finding is present, the *Attribute Agent* characterizes present findings in terms of severity and location, and the *Temporal Change Agent* identifies how each finding has changed since the prior study. We further perform verification at two stages. Before report generation, the *Consistency Gate* corrects change states that are incompatible with the diagnosis state. After report generation, the *Validation Agent* checks whether the generated report is supported by the structured clinical state. Figure 2 illustrates the overall pipeline.

Clinical Decision Agents

Diagnosis Agent Let $\mathcal{D}$ denote the set of 14 CheXpert findings, consisting of 12 disease-related findings and two non-disease labels, $\{\textit{Support Devices}, \textit{No Finding}\}$. The Diagnosis Agent independently estimates the state of every finding $d \in \mathcal{D}$ from the corresponding study. It receives complementary evidence from seven frozen pretrained chest X-ray experts, comprising three classification experts and four generative experts. The classification experts provide continuous finding scores, whereas the generative experts produce report-like textual evidence that is converted into categorical finding states by a CheXbert labeler (Smit et al. 2020).

Let $\mathbf{e}_j$ denote the resulting expert evidence at time j. The expert evidence $\mathbf{e}_j$ and acquisition view v_j are serialized in a fixed finding order and integrated by an instruction-tuned LLM $\mathcal{A}_D$:

$$\hat{\mathbf{s}}_j = \mathcal{A}_D(\mathbf{e}_j, v_j) \in \mathcal{Y}_D^{|\mathcal{D}|}, \tag{2}$$

where $\mathcal{Y}_D = \{\texttt{POS}, \texttt{UNC}, \texttt{NEG}\}$ and $\hat{s}_{j,d}$ denotes the predicted state of finding d at time j.

Attribute Agent The Attribute Agent characterizes findings predicted as positive in the current study. For each such finding, a medical vision-language model $\mathcal{A}_A$ is prompted with the current chest X-ray I_t and returns its severity and location, where location may include laterality and anatomical region. We apply the Attribute Agent to the 12 disease-related findings, denoted by $\mathcal{D}_A = \mathcal{D} \setminus \{\textit{Support Devices}, \textit{No Finding}\}$.

For each $d \in \mathcal{D}_A$, the attribute output is defined as

$$\hat{\mathbf{a}}_{t,d} = \begin{cases} \mathcal{A}_A(I_t, d), & \hat{s}_{t,d} = \texttt{POS}, \\ \perp, & \text{otherwise}, \end{cases} \tag{3}$$

where $\mathcal{A}_A(I_t, d) = (\hat{a}^{\text{sev}}_{t,d}, \hat{a}^{\text{loc}}_{t,d})$ gives the predicted severity and location, and $\perp$ indicates that no attribute prediction is available because the finding is not diagnosed as positive.

Temporal Change Agent The Temporal Change Agent $\mathcal{A}_T$ determines how each finding has changed from the prior to the current study. It operates on the same 12 disease-related findings as the Attribute Agent, and we therefore define $\mathcal{D}_T = \mathcal{D}_A$. The agent produces a change-state

$$\tilde{\mathbf{m}}_t = \mathcal{A}_T\left(r_{t-1}, \delta_t, \hat{\mathbf{s}}_{t-1}, \hat{\mathbf{s}}_t, \mathbf{p}_{t-1}, \mathbf{p}_t\right), \tag{4}$$

where r_{t-1} is the prior report, δ_t is the inter-study interval, and $\hat{\mathbf{s}}_{t-1}$ and $\hat{\mathbf{s}}_t$ are the diagnosis states of the prior and current studies, respectively. The $\mathbf{p}_{t-1}$ and $\mathbf{p}_t$, produced by the three classification experts used in the Diagnosis Agent, contain the predicted presence probabilities for the 12 disease-related findings in the prior and current studies, respectively. Each element $\tilde{m}_{t,d}$ represents the change state of finding d and takes one of six labels $\mathcal{Y}_T = \{\texttt{new}, \texttt{increased}, \texttt{stable}, \texttt{decreased}, \texttt{resolved}, \texttt{none}\}$. The label `none` indicates that no longitudinal change state is assigned to the finding, so that a change state is defined for every finding in $\mathcal{D}_T$. Rather than directly comparing the raw image pair, the agent uses disease-specific presence probabilities and the prior report, providing an explicit and interpretable basis for finding-wise temporal reasoning.

Training Strategies for Clinical Decision Agents

Supervised Fine-Tuning We first optimize each of the three agents using supervised fine-tuning (SFT). The Diagnosis Agent $\mathcal{A}_D$ is fine-tuned to predict the finding states in $\mathcal{Y}_D$, while the Attribute Agent $\mathcal{A}_A$ is trained to estimate the severity and location of positive findings. The attribute targets are extracted from the training reports using an LLM-based extractor. The Temporal Change Agent $\mathcal{A}_T$ is fine-tuned to predict disease-wise change states defined in $\mathcal{Y}_T$. To promote consistency across both temporal directions, we augment each valid training pair with a time-reversed counterpart. The prior and current evidence are swapped, along with the corresponding change labels: `new` and `resolved` are interchanged, as are `increased` and `decreased`, while `stable` and `none` remain unchanged. This augmentation encourages directionally consistent predictions when the study order is reversed.

Progression-Aware GRPO for Temporal Change Agent We post-train the Temporal Change Agent $\mathcal{A}_T$ with a two-stage pipeline: SFT warmup followed by a reinforcement-learning (RL) stage (Guo et al. 2025). In the RL stage we optimize $\mathcal{A}_T$ with Group Relative Policy Optimization (GRPO) (Shao et al. 2024). Since the reward is computed programmatically from the structured change labels rather than by a learned reward model, our objective is an instance of RLVR. For each input, the policy samples a group of G candidate responses $\{y_i\}_{i=1}^{G}$. Each response is assigned a reward $R_i = R(y_i, y^*)$ with respect to the report-derived target y^*. The policy is then updated using the standard clipped GRPO objective, in which the advantage of y_i is its reward standardized within the group.

The reward is a shaped reward designed to reflect the structured relationships among longitudinal change states. A naive outcome (exact-match) reward treats all incorrect predictions equally, whereas our multi-level formulation assigns partial credit to direction-preserving errors. For example, predicting `increased` when the target is `new` preserves the worsening direction, whereas predicting `decreased` indicates the opposite direction. We therefore evaluate each response at three levels: detection, coarse-grained direction, and fine-grained change state.

$$R(y, y^*) = \frac{1}{3}\left(\mathrm{F1}_{\mathrm{detect}} + \mathrm{F1}_{\mathrm{coarse}} + \mathrm{F1}_{\mathrm{fine}}\right). \quad (5)$$

At the detection level, the five explicit change states are grouped together, while `none` indicates that no change state is assigned. At the coarse level, `new` and `increased` are grouped as worsening, `decreased` and `resolved` as improving, while `stable` and `none` remain separate. At the fine level, predictions are evaluated using the original six change-state labels. This multi-level reward gives partial credit to predictions that preserve the clinical direction but miss the fine-grained change state, while assigning lower rewards to direction reversals and omitted changes.

Verified Structured Clinical State

Consistency Gate Role-specialized clinical decision agents may yield mutually inconsistent predictions, as each agent addresses a distinct clinical objective. We therefore introduce a deterministic Consistency Gate that enforces two coherence properties between the diagnosis and change states, leaving all other predictions unchanged. First, a diagnosis of `NEG` is incompatible with `new` and `increased`, which the gate maps to `decreased` if the finding is present in the prior report and to `none` otherwise, whereas a diagnosis of `POS` is incompatible with `resolved`, which is replaced by `decreased` to preserve the direction of change. Second, a finding diagnosed as `POS` must carry a temporal status, since the five change states are exhaustive for a present finding; the gate therefore completes a remaining `none` to `stable`, the only state that asserts no change. The corrected change state is denoted by $\bar{m}_{t,d}$.

Structured Clinical State We aggregate the outputs of the clinical decision agents after consistency correction into a Structured Clinical State (SCS), denoted by $\mathcal{S}_t$. For each finding d, we define

$$\mathbf{z}_{t,d} = (\hat{s}_{t,d}, \hat{\mathbf{a}}_{t,d}, \bar{m}_{t,d}), \qquad \mathcal{S}_t = \{\mathbf{z}_{t,d}\}_{d\in\mathcal{D}}. \quad (6)$$

Here, $\hat{\mathbf{a}}_{t,d} = \perp$ when attribute estimation is unavailable or not applicable, whereas $\bar{m}_{t,d} = \texttt{none}$ indicates that no explicit temporal change is assigned. The SCS preserves explicit disease-wise evidence for report generation and supports tracing errors to the corresponding clinical decision.

State-Grounded Report Generation and Validation

Report Generation A retrieval module selects the top-K training reports using IDF-weighted cosine similarity between binary disease signatures derived from $\mathcal{S}_t$, yielding the exemplar set $\mathcal{R}_t$. A frozen image-to-report model $\mathcal{B}$ generates a current-study draft b_t, and a frozen instruction-tuned Writer $\mathcal{W}$ produces the report conditioned on the SCS, the base draft, the retrieved exemplars, and the clinical context:

$$b_t = \mathcal{B}(X_t, c_t), \qquad r_t^{\mathrm{W}} = \mathcal{W}(\mathcal{S}_t, b_t, \mathcal{R}_t, c_t). \quad (7)$$

The SCS provides the primary clinical content, while the base draft and retrieved reports supply complementary image details and stylistic guidance.

Input	Method	Venue	NLG Metrics ↑						CE Metrics ↑		
			B-1	B-2	B-3	B-4	R-L	MTR	P	R	F1
Single Image	R2Gen	EMNLP'20	0.308	0.190	0.126	0.089	0.266	0.124	0.457	0.290	0.355
	R2GenCMN	ACL'21	0.328	0.203	0.135	0.096	0.273	0.133	0.512	0.359	0.422
	R2GenGPT	Meta-Rad'23	0.396	0.243	0.161	0.108	0.260	0.155	0.495	0.385	0.404
	PromptMRG	AAAI'24	0.384	0.229	0.149	0.104	0.261	0.146	0.517	0.453	0.439
	EKAGen	CVPR'24	0.405	0.245	0.156	0.105	0.264	0.154	0.472	0.404	0.411
	GMoD	MICCAI'24	0.378	0.234	0.155	0.107	0.276	0.162	0.496	0.429	0.460
	RADAR	ACL'25	0.412	0.242	0.162	0.114	0.257	0.155	0.448	0.436	0.417
	MedRAX	ICML'25	0.388	0.235	0.151	0.102	0.259	0.151	0.524	0.507	0.515
Longitudinal Images	Prefilling	MICCAI'23	0.343	0.210	0.141	0.100	0.274	0.137	0.506	0.364	0.423
	HERGen	ECCV'24	0.389	0.242	0.163	0.117	0.282	0.155	0.421	0.289	0.295
	STREAM	TMI'25	0.394	0.237	0.144	0.104	0.261	0.143	0.472	0.428	0.411
	MLRG	CVPR'25	0.416	0.252	0.157	0.114	0.264	0.158	0.507	0.425	0.418
	LLM-RG4	AAAI'25	0.417	0.240	0.155	0.115	0.257	0.147	0.498	0.441	0.436
	HC-LLM	AAAI'25	0.404	0.247	0.164	0.116	0.271	0.163	0.488	0.415	0.448
	Diff-RRG	MICCAI'25	0.405	0.251	0.169	0.120	0.276	0.164	0.528	0.430	0.474
	PriorRG	AAAI'26	0.369	0.261	<u>0.199</u>	<u>0.159</u>	**0.337**	0.170	<u>0.576</u>	0.452	0.507
	MARE	AAAI'26	0.409	<u>0.265</u>	0.184	0.133	0.291	0.161	0.433	0.378	0.375
	BiOTPrompt	CVPR'26	0.397	0.253	0.174	0.126	0.285	0.155	0.471	0.424	0.417
	TIM	CVPR'26	<u>0.430</u>	<u>0.265</u>	0.179	0.124	0.287	<u>0.185</u>	0.563	<u>0.505</u>	<u>0.511</u>
	Ours	-	**0.466**	**0.318**	**0.235**	**0.183**	<u>0.335</u>	**0.195**	**0.581**	**0.665**	**0.620**

Table 1: Comparison with single-image and longitudinal RRG methods on Longitudinal-MIMIC. B-n, R-L, and MTR denote BLEU-n, ROUGE-L, and METEOR. P, R, and F1 are the micro-averaged CheXbert precision, recall, and F1. **Bold** and <u>underline</u> indicate the best and second-best result per metric.

Validation Agent Although the SCS explicitly specifies the clinical content to be reported, the generation process may still omit committed findings or express them inconsistently (Nishino et al. 2022). The Validation Agent therefore compares the draft report r_t^{W} with $\mathcal{S}_t$ and performs targeted local edits. It first extracts the diagnosis states from the draft using a CheXbert labeler. Positive findings missing from the report are added, whereas positive statements unsupported by the SCS are removed only when the diagnosis experts also provide weak current-image evidence. A second pass verifies that each change state is expressed for the correct finding and corrects missing or inconsistent descriptions. The resulting report is denoted by $\hat{r}_t$.

Experiments

Experimental Setup

Dataset We evaluate on Longitudinal-MIMIC (Zhu et al. 2023), the longitudinal subset of MIMIC-CXR (Johnson et al. 2024), and the standard benchmark for LRRG on which most recent longitudinal methods are compared (Dong et al. 2026; Gao et al. 2026; Yun et al. 2025). Each example pairs a current study with the patient's most recent prior study and its report, and the reference report accordingly describes both the current findings and how they have changed relative to the prior study. We follow the official train, validation, and test partitions, with $2{,}058$ studies in the test set, and evaluate against the released reference reports.

Implementation Details All three clinical decision agents are optimized through supervised fine-tuning, and the Temporal Change Agent is subsequently refined using the Progression-Aware GRPO objective in Eq. 5. Model checkpoints are selected according to validation performance. For report generation, a frozen PriorRG model (Liu et al. 2026a) provides a current-study draft, and we set $K{=}3$ for the retrieved exemplars. A frozen instruction-tuned 27B LLM performs both report writing and validation-based editing.

Baselines We compare against two groups of methods. Single-image RRG maps one chest X-ray to a report without temporal context and comprises R2Gen (Chen et al. 2020), R2GenCMN (Chen et al. 2021), R2GenGPT (Wang et al. 2023), PromptMRG (Jin et al. 2024), EKAGen (Bu et al. 2024), GMoD (Xiang et al. 2024), RADAR (Hou et al. 2025), and MedRAX (Fallahpour et al. 2025). Longitudinal RRG conditions on prior images, reports, or clinical context and comprises Prefilling (Zhu et al. 2023), HERGen (Wang, Du, and Yu 2024), STREAM (Yang et al. 2025a), MLRG (Liu et al. 2025a), LLM-RG4 (Wang et al. 2025), HC-LLM (Liu et al. 2025b), Diff-RRG (Yun et al. 2025), PriorRG (Liu et al. 2026a), MARE (Gao et al. 2026), BiOTPrompt (Liu et al. 2026b), and TIM (Dong et al. 2026).

For Table 1, we use the Longitudinal-MIMIC results reported in the original baseline papers when available and otherwise adopt the corresponding reimplementation results from TIM (Dong et al. 2026). If neither source provides the results, we train and evaluate the method using its official code. Table 2 requires access to the generated reports, so we retrain under the same protocol every baseline whose code is publicly released; MARE and TIM are therefore excluded, whereas MedRAX, although single-image, is retained as a representative agentic method.

Method	LCC ↑		ReXrank ↑						
	LCC-C	LCC-F	1/RadCliQ-v1	BLEU	BERTScore	SembScore	RadGraph-F1	RaTEScore	GREEN
Prefilling	0.089	0.053	0.819	0.176	0.397	0.352	0.180	0.528	0.281
HERGen	0.163	0.100	0.866	0.200	0.410	0.369	0.201	0.538	0.298
STREAM	0.166	0.100	0.820	0.195	0.401	0.333	0.191	0.537	0.277
MLRG	0.141	0.082	0.819	0.193	0.377	0.374	0.189	0.530	0.299
LLM-RG4	0.185	0.115	0.925	0.216	0.428	0.395	<u>0.216</u>	0.554	0.337
HC-LLM	0.155	0.089	0.873	0.214	0.421	0.356	0.205	0.540	0.297
Diff-RRG	0.187	0.114	0.900	0.220	0.426	0.376	0.211	0.542	0.322
PriorRG	0.187	0.115	<u>1.141</u>	<u>0.270</u>	**0.484**	<u>0.441</u>	**0.270**	<u>0.577</u>	<u>0.341</u>
BiOTPrompt	0.154	0.092	0.907	0.214	0.427	0.381	0.212	0.543	0.315
MedRAX	<u>0.193</u>	<u>0.128</u>	0.838	0.189	0.402	0.332	0.211	0.551	0.298
Ours (w/o Temporal)	0.148	0.095	1.132	0.267	0.467	0.475	0.260	0.582	0.337
Ours	**0.394**	**0.283**	**1.173**	**0.272**	<u>0.483</u>	**0.467**	**0.270**	**0.582**	**0.342**

Table 2: LCC and ReXrank results on Longitudinal-MIMIC. **Bold** and <u>underline</u> indicate the best and second-best results per metric. The ablation variant is excluded from the ranking.

Variant	B-1	MTR	CE-F1	LCC-C	LCC-F
Full model	0.466	0.195	0.620	0.394	0.283
w/o Attribute	0.454	0.191	0.615	0.346	0.250
w/o Temporal	0.452	0.192	0.622	0.148	0.095
w/o GRPO	0.464	0.195	0.618	0.389	0.259
w/o Consistency	0.465	0.195	0.620	0.389	0.274
w/o Base draft	0.381	0.171	0.610	0.404	0.280
w/o Validation	0.464	0.195	0.615	0.338	0.233
w/o Diagnosis	0.458	0.193	0.612	0.386	0.278

Table 3: Component ablation on Longitudinal-MIMIC.

Evaluation Metrics

Report Quality and Clinical Efficacy We report standard natural language generation (NLG) metrics, namely BLEU-1 to BLEU-4 (Papineni et al. 2002), ROUGE-L (Lin 2004), and METEOR (Banerjee and Lavie 2005). We measure clinical efficacy (CE) by using CheXbert (Smit et al. 2020) on the generated and the reference report and comparing the resulting 14-finding labels, reporting micro-averaged precision, recall, and F1. To assess clinical quality beyond surface overlap, we further adopt the ReXrank (Zhang et al. 2025) suite, which measures lexical and semantic similarity through BERTScore (Zhang et al. 2020) and SembScore (Smit et al. 2020) and clinically aware quality through RadGraph-F1 (Jain et al. 2021), $1/$RadCliQ-v1 (Yu et al. 2023), RaTEScore (Zhao et al. 2024), and GREEN (Ostmeier et al. 2024), together with a corpus-level BLEU.

Longitudinal Change Concordance To directly evaluate longitudinal correctness, we report LCC, which measures whether a generated report preserves the finding-specific temporal changes stated in the reference. An instruction-tuned Gemma-4-31B extracts change propositions, each consisting of a finding and one of five labels: `new`, `increased`, `decreased`, `resolved`, or `stable`. This label set omits `none`, since a proposition is extracted only where a report states a change. A deterministic scorer matches propositions from the generated and reference reports at two levels. Coarse matching (LCC-C) requires agreement on the finding and its clinical direction, collapsing `new` into `increased` and `resolved` into `decreased`, whereas fine matching (LCC-F) requires the exact change label. Macro-F1 is computed over the reference-stated changes at each level. Unlike general semantic and clinically aware report-level metrics, this reference-anchored evaluation penalizes omitted changes and distinguishes direction-preserving errors from direction reversals.

Main Results and Discussion

Comparison with the State of the Art Table 1 compares STRIVE with single-image and longitudinal RRG methods on Longitudinal-MIMIC. STRIVE achieves the best performance on all NLG metrics except ROUGE-L, where it ranks second, and on all three CE metrics. The larger gains in CE than in NLG suggest that the improvement primarily reflects clinical correctness rather than surface-level similarity, supported by explicitly representing diagnosis, attributes, and progression in the SCS before report generation. Moreover, recall exceeding precision (0.665 vs. 0.581) suggests that the framework recovers under-expressed clinical findings, while its leading precision indicates that this broader coverage does not introduce unsupported findings.

Table 2 evaluates report quality with the ReXrank suite, which measures lexical, semantic, and clinically aware agreement with the reference. STRIVE obtains the best result on six of its seven metrics, namely $1/$RadCliQ-v1 (1.173), BLEU (0.272), SembScore (0.467), RadGraph-F1 (0.270), RaTEScore (0.582), and GREEN (0.342), and ranks second on BERTScore. Since these metrics reward factual and relational agreement rather than n-gram overlap, organizing clinical evidence before generation improves not only what the report states but how faithfully it does so.

Longitudinal Correctness Table 2 evaluates longitudinal correctness using LCC. STRIVE achieves an LCC-C of 0.394 and an LCC-F of 0.283, more than twice the strongest baseline scores of 0.193 and 0.128, respectively. These gains in-

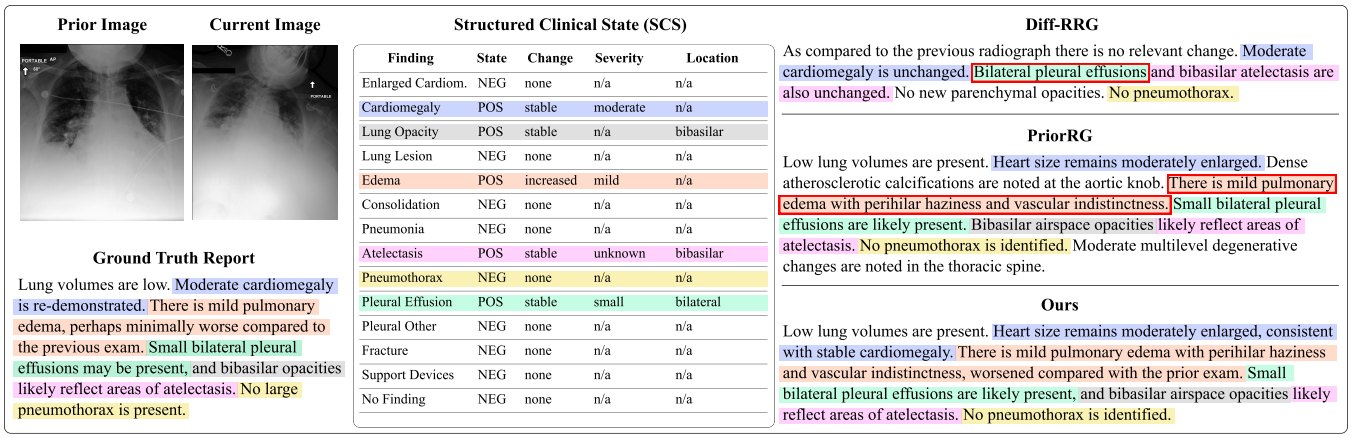

Figure 3: Qualitative comparison on Longitudinal-MIMIC. Colors indicate disease states matched to the reference report, while red boxes denote incomplete states. STRIVE captures all findings and their detailed clinical states.

dicate improved accuracy in both clinical direction and fine-grained change states, whereas all baselines remain below 0.20 despite competitive ReXrank performance. Removing the Temporal Change Agent collapses LCC while leaving ReXrank performance largely unchanged, showing that the LCC improvement is attributable to explicit temporal change reasoning rather than to general report quality.

Ablation Study

Table 3 presents ablation results on Longitudinal-MIMIC, where each variant removes or replaces a single component while keeping the remaining pipeline and decoding settings unchanged. The ablation results yield three insights. First, the Temporal Change Agent is the primary contributor to longitudinal correctness. Removing it sharply reduces LCC-C from 0.394 to 0.148 and LCC-F from 0.283 to 0.095, while CE-F1 remains essentially unchanged. Without the GRPO stage, LCC-F decreases to 0.259, indicating that the proposed reward improves fine-grained discrimination among change states. Second, the two verification stages provide complementary benefits. Removing the Validation Agent reduces CE-F1 from 0.620 to 0.615, LCC-C from 0.394 to 0.338, and LCC-F from 0.283 to 0.233, demonstrating its role in recovering findings and change states omitted during report generation. In contrast, removing the Consistency Gate leaves CE-F1 unchanged at 0.620 but lowers LCC-C from 0.394 to 0.389 and LCC-F from 0.283 to 0.274. This indicates that the gate specifically improves the consistency of longitudinal change predictions without altering finding-presence accuracy. Third, the Diagnosis and Attribute Agents, together with the base draft, contribute complementary information. Replacing the Diagnosis Agent with rule-based voting over the same seven experts reduces CE-F1 to 0.612, demonstrating the benefit of learned aggregation across heterogeneous experts. Removing the Attribute Agent lowers both NLG and CE performance, indicating the importance of severity and location information. In contrast, substituting the prior report for the base draft causes the largest NLG decline, with BLEU-1 decreasing from 0.466 to 0.381, but only a modest reduction in CE-F1. This suggests that the base draft primarily supports report realization, while the clinical decision agents determine the core clinical content.

Qualitative Analysis

Figure 3 compares STRIVE with Diff-RRG and PriorRG on a representative longitudinal case, together with the SCS committed before generation. Both baselines recognize several findings from the reference report but leave their clinical state incomplete. Diff-RRG omits edema and lung opacity and reports pleural effusion without its severity, whereas PriorRG recognizes the pulmonary edema but not its interval increase. The SCS instead captures every finding stated in the reference together with its severity, location, and change, and STRIVE realizes all of these fields in the final report. The generated report therefore agrees with the reference not only on which abnormalities are present but also on how they are characterized and how they have changed.

Conclusion

In this work, we proposed STRIVE, a multi-agent framework for LRRG. Rather than handling all clinical reasoning tasks within a single model, it assigns diagnosis, clinical attribute estimation, and temporal change reasoning to specialized agents. Furthermore, Progression-Aware GRPO improves directional change modeling, while the Consistency Gate and Validation Agent ensure consistency between structured clinical evidence and the generated report. On Longitudinal-MIMIC, STRIVE improves report quality, clinical efficacy, and LCC, more than doubling the LCC of the strongest baseline.

References

Banerjee, S.; and Lavie, A. 2005. 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*, 65–72.

Bannur, S.; Bouzid, K.; Castro, D. C.; Schwaighofer, A.; Thieme, A.; Bond-Taylor, S.; Ilse, M.; Pérez-García, F.; Salvatelli, V.; Sharma, H.; et al. 2024. MAIRA-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Brady, A. P. 2018. Radiology reporting—from Hemingway to HAL? *Insights into Imaging*, 9(2): 237–246.

Bu, S.; Li, T.; Yang, Y.; and Dai, Z. 2024. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14194–14204.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021. 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)*, 5904–5914.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 1439–1449.

Chen, Z.; Varma, M.; Delbrouck, J.-B.; Paschali, M.; Blankemeier, L.; Van Veen, D.; Valanarasu, J. M. J.; Youssef, A.; Cohen, J. P.; Reis, E. P.; et al. 2024. CheXagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Dong, Y.; Lin, Y.; Huang, S.; Yang, X.; and Yang, X. 2026. 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*, 6951–6961.

Fallahpour, A.; Ma, J.; Munim, A.; Lyu, H.; and Wang, B. 2025. MedRAX: Medical Reasoning Agent for Chest X-ray. In *International Conference on Machine Learning*, 15661–15676. PMLR.

Gao, Q.; Liu, T.; Li, X.; Zhang, X.; Sun, Z.; Wang, B.; Yin, B.; and Liu, Z. 2026. MARE: Multimodal analogical reasoning for disease evolution-aware radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 21180–21188.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hou, W.; Cheng, Y.; Xu, K.; Li, H.; Hu, Y.; Li, W.; and Liu, J. 2025. RADAR: Enhancing radiology report generation with supplementary knowledge injection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 26366–26381.

Jain, S.; Agrawal, A.; Saporta, A.; Truong, S.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M. P.; Ng, A. Y.; Langlotz, C.; et al. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Jin, H.; Che, H.; Lin, Y.; and Chen, H. 2024. PromptMRG: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2607–2615.

Johnson, A.; Lungren, M.; Peng, Y.; Lu, Z.; Mark, R.; Berkowitz, S.; and Horng, S. 2024. MIMIC-CXR-JPG - chest radiographs with structured labels (version 2.1.0).

Khan, S. U. R.; Maqsood, H.; Vollmer, S.; Dengel, A.; and Asim, M. N. 2026. CogRad: A Cognitively-Inspired Multi-Agent Framework for Radiology Report Generation. *arXiv preprint arXiv:2607.03853*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81.

Liu, K.; Ma, Z.; Fang, Z.; Li, Y.; Xie, K.; and Miao, Q. 2026a. PriorRG: Prior-guided contrastive pre-training and coarse-to-fine decoding for chest x-ray report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 7206–7214.

Liu, K.; Ma, Z.; Kang, X.; Li, Y.; Xie, K.; Jiao, Z.; and Miao, Q. 2025a. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10348–10359.

Liu, R.; Li, M.; Zhao, S.; Chen, L.; Chang, X.; and Yao, L. 2024. In-context learning for zero-shot medical report generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8721–8730.

Liu, T.; Fan, Y.; Wang, B.; Hu, Y.; Li, M.; Li, J.; and Gao, J. 2026b. 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 (CVPR)*, 612.

Liu, T.; Wang, J.; Hu, Y.; Li, M.; Yi, J.; Chang, X.; Gao, J.; and Yin, B. 2025b. HC-LLM: Historical-constrained large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 5595–5603.

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11976–11986.

Nicolson, A.; Dowling, J.; Anderson, D.; and Koopman, B. 2024. Longitudinal data and a semantic similarity reward for chest X-ray report generation. *Informatics in Medicine Unlocked*, 50: 101585.

Nishino, T.; Miura, Y.; Taniguchi, T.; Ohkuma, T.; Suzuki, Y.; Kido, S.; and Tomiyama, N. 2022. Factual Accuracy is not Enough: Planning Consistent Description Order for Radiology Report Generation. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7123–7138. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Ostmeier, S.; Xu, J.; Chen, Z.; Varma, M.; Blankemeier, L.; Bluethgen, C.; Md, A.; Moseley, M.; Langlotz, C.; Chaudhari, A.; et al. 2024. GREEN: Generative Radiology Report

Evaluation and Error Notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 374–390.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pérez-García, F.; Sharma, H.; Bond-Taylor, S.; Bouzid, K.; Salvatelli, V.; Ilse, M.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Lungren, M. P.; et al. 2025. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 1–12.

Sellergren, A.; Kazemzadeh, S.; Jaroensri, T.; Kiraly, A.; Traverse, M.; Kohlberger, T.; Xu, S.; Jamil, F.; Hughes, C.; Lau, C.; et al. 2025. MedGemma technical report. *arXiv preprint arXiv:2507.05201*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A. Y.; and Lungren, M. 2020. CheXbert: 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*, 1500–1519.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7433–7442.

Wang, F.; Du, S.; and Yu, L. 2024. HERGen: Elevating radiology report generation with longitudinal data. In *European Conference on Computer Vision*, 183–200. Springer.

Wang, X.; Li, Y.; Wang, F.; Wang, S.; Li, C.; and Jiang, B. 2024. R2GenCSR: Retrieving Context Samples for Large Language Model based X-ray Medical Report Generation. *CoRR*.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023. R2GenGPT: Radiology report generation with frozen LLMs. *Meta-Radiology*, 1: 100033.

Wang, Z.; Sun, Y.; Li, Z.; Yang, X.; Chen, F.; and Liao, H. 2025. LLM-RG4: Flexible and factual radiology report generation across diverse input contexts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8250–8258.

Xiang, Z.; Cui, S.; Shang, C.; Jiang, J.; and Zhang, L. 2024. GMoD: graph-driven momentum distillation framework with active perception of disease severity for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 295–305. Springer.

Yang, Y.; You, X.; Zhang, K.; Fu, Z.; Wang, X.; Ding, J.; Sun, J.; Yu, Z.; Huang, Q.; Han, W.; et al. 2025a. Spatio-temporal and retrieval-augmented modelling for chest X-ray report generation. *IEEE Transactions on Medical Imaging*.

Yang, Z.; Xu, X.; Zhang, J.; Wang, G.; Kalra, M. K.; and Yan, P. 2025b. Chest X-ray foundation model with global and local representations integration. *IEEE Transactions on Medical Imaging*.

Yu, F.; Endo, M.; Krishnan, R.; Pan, I.; Tsai, A.; Reis, E. P.; Fonseca, E. K. U. N.; Lee, H. M. H.; Abad, Z. S. H.; Ng, A. Y.; et al. 2023. Evaluating progress in automatic chest X-ray radiology report generation. *Patterns*, 4(9).

Yun, H.; Maeng, J.; Kang, E.; and Suk, H.-I. 2025. 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*, 152–161. Springer.

Zhang, E.; Hao, Y.; Wang, L.; and Guo, Z. 2026a. The Double Dilemma in Multi-Task Radiology Report Generation: A Gradient Dynamics Analysis and Solution. In *Forty-third International Conference on Machine Learning*.

Zhang, K.; Barrett, C. D.; Kim, J.; Sun, L.; Taghavi, T.; and Kenthapadi, K. 2026b. RadAgents: Multimodal Agentic Reasoning for Chest X-ray Interpretation with Radiologist-like Workflows. In *Proceedings of the 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, 3496–3519. PMLR.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zhang, X.; Zhou, H.-Y.; Yang, X.; Banerjee, O.; Acosta, J. N.; Miller, J.; Huang, O.; and Rajpurkar, P. 2025. ReXrank: A Public Leaderboard for AI-Powered Radiology Report Generation. In *AAAI Bridge Program on AI for Medicine and Healthcare*, 90–99. PMLR.

Zhao, W.; Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. RaTEScore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15004–15019.

Zhu, Q.; Mathai, T. S.; Mukherjee, P.; Peng, Y.; Summers, R. M.; and Lu, Z. 2023. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 189–198. Springer.

Supplementary Material

S1 Dataset

S1.1 Dataset Construction

Longitudinal-MIMIC (Zhu et al. 2023) is built from the 26,625 MIMIC-CXR (Johnson et al. 2024) patients with two visits. Studies whose report has no findings section are discarded, and the remainder is partitioned with the official MIMIC-CXR patient-level split, so no patient appears in more than one partition. Each example pairs the current study with the patient's most recent prior study and its report. Table S1 lists the resulting statistics.

S1.2 Temporal Change Statistics

To characterize how much temporal content the reports actually carry, we apply the LCC change-statement extractor of Section S4 to the reference reports of all three partitions. Two properties motivate the LCC evaluation. First, temporal statements are common but far from universal: a comparison is stated in 55.0% of the training reports, 56.2% of the validation reports, and 67.5% of the test reports, at a density of 1.27 to 1.59 statements per report. Second, the change labels are strongly imbalanced in every partition. On the test partition they are dominated by `stable` (2,219), followed by `increased` (410), `decreased` (384), `new` (175), and `resolved` (87), so the rarest state accounts for under 3% of the statements. We therefore compute LCC as a macro-F1 over the change labels, so that the rare directional states are not absorbed by the dominant `stable` class.

S2 Implementation Details

S2.1 Clinical Decision Agents

Training Configuration The Diagnosis and Temporal Change Agents use Gemma-4-E4B-it; the Attribute Agent uses MedGemma-1.5-4B-it. All backbones are frozen and adapted with completion-only bfloat16 LoRA (r=16, α=32, dropout 0.05). Checkpoints are selected on the patient-disjoint validation partition. Table S2 gives the optimization settings. Training uses two NVIDIA RTX A6000 GPUs with 48 GB each.

Chest X-ray Expert Pool for the Diagnosis Agent The Diagnosis Agent aggregates three classification experts (ConvNeXt (Liu et al. 2022), RAD-DINO (Pérez-García et al. 2025), and CheXFound (Yang et al. 2025b)) and four generative experts (MedGemma (Sellergren et al. 2025), PriorRG (Liu et al. 2026a), CheXagent (Chen et al. 2024), and MAIRA-2 (Bannur et al. 2024)). All seven experts are frozen while the Diagnosis Agent is trained and evaluated. The classification experts predict continuous scores for the 14 CheXpert findings, which also provide the per-finding probabilities $\mathbf{p}_{t-1}$ and $\mathbf{p}_t$ for the Temporal Change Agent. A frozen CheXbert labeler (Smit et al. 2020) maps the reports of the generative experts to categorical finding states, and all evidence is serialized in a fixed finding order.

	Train	Val	Test
Patients	26,156	203	266
Studies	92,374	737	2,058
Reports with change statements	50,849	414	1,390
Finding-level change statements	117,040	954	3,275

Table S1: Longitudinal-MIMIC partitions. Each study is paired with the patient's most recent prior study, and the data are split at the patient level to prevent patient overlap across partitions.

	Diag.	Attr.	Temporal	
	SFT	SFT	SFT	GRPO
Backbone	Gemma	MedGemma	Gemma	SFT ckpt
LoRA r/α	16/32	16/32	16/32	16/32
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	$1e^{-4}$	$1e^{-4}$	$1e^{-4}$	$1e^{-5}$
Batch size	4	4	4	4
Precision	bf16	bf16	bf16	bf16

Table S2: Optimization settings of the clinical decision agents. GRPO is initialized from the corresponding SFT checkpoint.

Progression-Aware GRPO for Temporal Change Agent The Temporal Change Agent is first trained with SFT and subsequently optimized with GRPO using the same LoRA configuration. For each prompt, the policy samples a group of G=6 completions at temperature 1.0. Group-relative advantages are computed using the progression-aware reward defined in the main paper. Because the reward is computed directly from the sampled change states and the report-derived target, GRPO does not require a separately learned reward model.

S2.2 Report Generation

Base Draft Generation The base draft b_t is produced by a frozen PriorRG (Liu et al. 2026a) image-to-report model, which supplies current-image detail such as devices and measurements. For retrieval, the positive findings of the Structured Clinical State (SCS) are converted into a binary CheXpert-14 signature and matched against the signatures of the training reports under an IDF-weighted cosine similarity, where the inverse document frequency of finding d is $\log\big((N{+}1)/(n_d{+}1)\big)+1$ for a corpus of N reports containing finding d in n_d of them. Weighting by the square root of the IDF on both sides makes the similarity favor exemplars whose rare findings match and penalize exemplars that carry extra findings. The top K=3 reports are injected as style exemplars.

Writer The Writer is a frozen instruction-tuned Qwen3.6-27B decoded greedily. Its prompt presents the SCS as the authoritative clinical content, the base draft as a source of image detail, and the retrieved reports as style references

only, and it instructs the model to resolve any conflict in favor of the SCS.

S2.3 Validation Agent

The Validation Agent uses the same frozen 27B model and greedy decoding as the Writer. Given a draft report, it first applies the frozen CheXbert labeler and compares the resulting 14-finding state vector with the SCS. Findings marked as positive in the SCS but omitted from the draft are inserted. Conversely, positive statements unsupported by the SCS are removed only when the diagnosis experts assign the corresponding finding a current-image probability below 0.3, thereby preventing deletion based solely on discrepancies in wording. In a second pass, the agent verifies whether each finalized temporal change state is expressed for the correct finding and locally revises the corresponding sentence when an inconsistency is detected, rather than regenerating the entire report. The validation procedure is limited to two passes.

S3 Evaluation Metrics

S3.1 NLG Metrics

These three metrics measure surface agreement with the reference wording.

- **BLEU-n (B-n)** (Papineni et al. 2002): the geometric mean of modified n-gram precisions up to order n, with a brevity penalty. It is a corpus-level statistic, since the n-gram counts of all $2{,}058$ studies are pooled before the precision is formed, and being precision-based it penalizes content the reference does not contain.
- **ROUGE-L (R-L)** (Lin 2004): the F-measure over the longest common subsequence of the two reports. It does not require contiguity, so it is the more recall-oriented of the two.
- **METEOR (MTR)** (Banerjee and Lavie 2005): a recall-weighted harmonic mean of unigram precision and recall over an alignment that admits stem, synonym, and paraphrase matches, with a penalty for fragmented alignments.

These overlap metrics do not distinguish clinically decisive terms from stylistic words.

S3.2 CE Metrics

CE metrics score what the report asserts rather than how it is worded.

- **Precision (P), Recall (R), and F1** are computed by labeling the generated and the reference report with the CheXbert labeler (Smit et al. 2020), which assigns each of 14 thoracic findings one of *present*, *absent*, *uncertain*, or *blank*. Following the convention of the baselines we compare against, the label is binarized by treating *present* as positive and the other three as negative, and the scores are micro-averaged over all finding–study pairs.

CE measures current-study finding presence, not temporal direction.

S3.3 ReXrank Metrics

The ReXrank suite (Zhang et al. 2025) contributes seven metrics that reach beyond the fourteen-label constraint of CheXbert.

- **BLEU-2**: the mean of the CXR-Report-Metric study-level bigram BLEU scores, computed with the suite's own preprocessing and reported separately from the corpus BLEU of Section S3.1.
- **BERTScore** (Zhang et al. 2020): an F-measure over greedily matched contextual token embeddings, so it credits equivalent phrasing that shares no surface tokens.
- **SembScore** (Smit et al. 2020): the cosine similarity between the CheXbert label embeddings of the two reports, an embedding-space relaxation of the CE comparison.
- **RadGraph-F1** (Jain et al. 2021): the F1 over the clinical entities and relations RadGraph extracts from each report.
- $1/$**RadCliQ-v1** (Yu et al. 2023): the reciprocal of a composite error score fitted by regression to radiologist error counts, reported as a reciprocal so that larger is better.
- **RaTEScore** (Zhao et al. 2024): an entity-level similarity that weights medical entities by clinical importance and is robust to synonymy and negation.
- **GREEN** (Ostmeier et al. 2024): a score derived from the clinically significant errors a language model enumerates between the two reports.

S3.4 Why LCC Is Needed

The report-level metrics above compare each generated report with the reference as a whole, allowing correctly generated findings and common report content to outweigh errors in a small number of temporal expressions. However, these expressions encode the key clinical information that distinguishes longitudinal reporting from single-study description. For example, pleural effusion has increased and pleural effusion remains stable identify the same current finding and share most of their wording, yet describe different disease trajectories. A more consequential error occurs when improved is replaced by worsened, reversing the clinical interpretation despite otherwise high report-level similarity. Temporal statements should therefore be evaluated according to their clinical significance rather than their relatively small lexical footprint.

The ablation results in the main paper support this distinction. Removing the Temporal Change Agent has little effect on ReXrank but substantially reduces LCC, indicating that whole-report similarity can remain high even when temporal reasoning deteriorates. LCC addresses this limitation by directly evaluating the temporal changes stated in the reference reports and distinguishing direction-preserving errors from reversals.

S4 Longitudinal Change Concordance

S4.1 Change-statement Extraction

LCC employs an instruction-tuned Gemma-4-31B to transform each report into a set of structured, finding-level

change statements. The extracted statements are subsequently matched and aggregated using deterministic rules. The model is used solely for extracting finding-level evidence and change labels and does not directly assign the metric score. To ensure consistent processing, the same model, prompt template, and decoding configuration are applied to both the reference reports and all generated reports.

The extractor is applied once to each report using the prompt template provided below. The corresponding user message specifies the 13 admissible CheXpert findings (all 14 categories except *No Finding*), the five admissible change labels, and the report text. Greedy decoding is used throughout. Prior to extraction, each report is screened using a deterministic pattern matcher; reports without comparison cues are assigned an empty statement list without invoking the LLM. The initial generation budget is 1,024 tokens. If decoding terminates at the length limit, the budget is iteratively doubled up to 4,096 tokens to prevent long reports containing multiple findings from being truncated into incomplete statement lists.

```
[SYSTEM]
You extract longitudinal change propositions
 from chest radiology reports.

Return compact JSON only: a list of objects.
 Each object must have exactly these keys:
- finding: one of the allowed findings
- label: one of new, increased, decreased,
resolved, stable
- evidence_span: the exact sentence or
clause supporting the proposition
- change_cue: the exact temporal/change
phrase
- mention: the finding phrase
- laterality: left, right, bilateral, or
empty string
- region: upper, mid, lower, hilar,
mediastinal, or empty string

Rules:
- Extract only explicit comparison/
progression statements.
- Do not infer change from current findings
alone.
- stable includes unchanged, stable, similar
, persistent, redemonstrated, no new, no
significant change.
- new means newly present/interval
development.
- increased means worsened/progressed/larger
/more extensive.
- decreased means improved/smaller/less
extensive/resolving/partial clearing.
- resolved means absent now after previously
 present/no longer seen/cleared.
- Ignore non-clinical dates and comparison
boilerplate unless a finding change is
stated.
- If no explicit finding-level change exists
, return [].
- Do not include explanations, markdown,
confidence, or extra keys.

[USER]
Allowed findings: Atelectasis, Cardiomegaly,
 Enlarged Cardiomediastinum, Consolidation,
Edema,
Lung Lesion, Lung Opacity, Pleural Effusion,
 Pleural Other, Pneumonia, Pneumothorax,
Fracture,
Support Devices
Allowed labels: new, increased, decreased,
resolved, stable

Report:
<report text>

JSON:
```

S4.2 Matching Rules

LCC uses reference-anchored, one-to-one matching. For each statement in the reference report, it searches the generated report for an unmatched candidate statement referring to the same finding. Laterality and anatomical region are retained by the extractor but are not imposed as hard matching constraints, because LCC is designed to isolate temporal agreement from localization specificity. When multiple compatible candidates are available, they are prioritized by exact fine-grained label agreement, followed by coarse directional agreement and then finding agreement alone. Each candidate statement can be matched at most once, preventing a single generated statement from covering multiple reference changes. If no compatible candidate is found, the reference statement is assigned the predicted label `none`, thereby encoding the omission as a false negative.

This produces one predicted label per reference statement, and macro-F1 is computed over those pairs, at the fine level over the five labels and at the coarse level after collapsing `new` into `increased` and `resolved` into `decreased`. *Support Devices* statements are excluded from the primary scope. Candidate-only statements are excluded from this reference-anchored primary score.

S5 In-Depth Analysis

S5.1 Longitudinal Change Coverage

Figure S1 compares omission rates to assess how well each method preserves the temporal changes stated in the reference reports. The baselines omit 0.71–0.87 of the reference changes, whereas STRIVE reduces the omission rate to 0.46, corresponding to an absolute reduction of 0.25 and a relative reduction of 35% over the best baseline. When the Temporal Change Agent is removed, the omission rate increases from 0.46 to 0.83, returning to the baseline range.

These results show that STRIVE's LCC gains arise not only from predicting the correct direction for mentioned changes but also from preserving substantially more reference-stated temporal information. Together with the improvements in coarse- and fine-grained LCC, the ablation confirms that the explicit finding-wise transition map improves both change coverage and directional accuracy by

Method	Fine change label (F1) ↑					LCC-F	Direction (F1) ↑		LCC-C
	new	increased	decreased	resolved	stable		worsened	improved	
Prefilling	0.000	0.024	0.005	0.000	0.236	0.053	0.027	0.004	0.089
HERGen	0.040	0.071	0.064	0.000	0.325	0.100	0.107	0.057	0.163
STREAM	0.039	0.086	0.039	0.000	0.336	0.100	0.127	0.036	0.166
MLRG	0.000	0.068	0.039	0.000	0.300	0.082	0.077	0.045	0.141
LLM-RG4	0.000	0.057	0.130	0.023	0.368	0.115	0.066	0.121	0.185
HC-LLM	0.000	0.052	0.076	0.000	0.318	0.089	0.049	0.099	0.155
Diff-RRG	0.000	0.097	0.066	0.000	0.405	0.114	0.083	0.074	0.187
PriorRG	0.050	0.068	0.084	0.000	0.373	0.115	0.105	0.085	0.187
BiOTPrompt	0.000	0.048	0.035	0.000	0.376	0.092	0.054	0.033	0.154
MedRAX	0.102	0.113	0.098	0.000	0.326	0.128	0.145	0.108	0.193
Ours	**0.195**	**0.316**	**0.200**	**0.193**	**0.511**	**0.283**	**0.407**	**0.265**	**0.394**
Reference n	175	410	384	87	2,219	3,275	585	471	3,275

Table S3: Per-label decomposition of LCC. Each cell is the F1 with which a system recovers the reference changes carrying that label; LCC-F and LCC-C are the macro means over the five fine labels and the three direction classes. The `stable` column is shared by both levels and is therefore not repeated in the direction block. The last row gives the number of reference changes carrying each label, which is the denominator of the corresponding recall. **Bold** and underline mark the best and second-best system per column.

providing the Writer with committed change states that the Validation Agent can enforce.

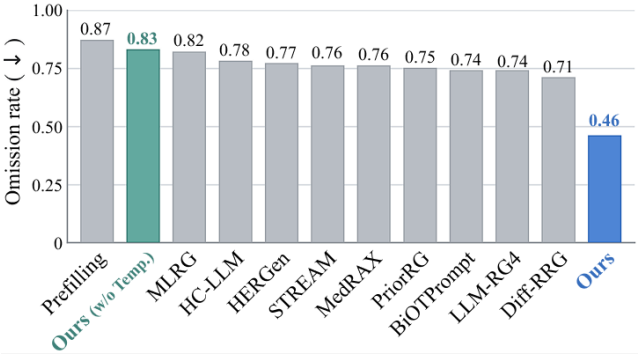

Figure S1: Omission rate under the primary direction-only matching protocol, defined as the fraction of reference-stated changes absent from the generated report. STRIVE achieves the lowest omission rate, while removing the Temporal Change Agent increases it to the baseline range.

S5.2 Per-Label Analysis of LCC

Table S3 decomposes LCC into per-label F1 scores, averaged over five fine-grained labels for LCC-F and three coarse-grained labels for LCC-C. STRIVE achieves the highest F1 for all five labels, confirming that its improvement is not driven solely by the dominant `stable` class. The largest gains are observed for `increased` and `resolved`; notably, nine of the ten baselines fail to recover any of the 87 reference `resolved` statements correctly. This result also motivates macro averaging: because `stable` accounts for 2,219 of the 3,275 reference changes, micro averaging would obscure performance on the rarer directional states, whereas macro averaging assigns equal weight to each label. This distinction is clinically important because reliable longitudinal reporting requires not only recognizing unchanged findings but also accurately capturing the rarer directional changes that indicate progression or resolution.

S5.3 Bootstrap Analysis of LCC Improvements

We assessed the statistical significance of the LCC improvements using a paired nonparametric bootstrap over the 2,058 test study pairs with 2,000 resamples. LCC was recomputed for each resample because it is a corpus-level macro-F1 rather than an average of study-level scores. As reported in Table S4, STRIVE's 95% confidence intervals lie entirely above those of all ten baselines at both granularities; for LCC-C, the strongest baseline has an upper bound of 0.212, whereas STRIVE has a lower bound of 0.373. The paired differences between STRIVE and every baseline remain positive across all resamples. Even against the strongest baseline, MedRAX, STRIVE achieves margins of 0.201 in LCC-C (95% CI $[0.174, 0.229]$) and 0.155 in LCC-F (95% CI $[0.126, 0.183]$), confirming that the improvements are statistically significant.

S5.4 Stage-Wise Temporal Information Analysis

Table S5 traces how temporal information is preserved across the pipeline by evaluating each stage against the same 3,275 reference change statements. The Temporal Change Agent achieves 0.342/0.241 LCC-C/F, which increases to 0.354/0.250 after the Consistency Gate. The Writer output decreases to 0.338/0.233, indicating that some structured change states are not verbalized, whereas Validation raises the final report to 0.394/0.283. The first two rows evaluate structured states and the last two evaluate report-extracted statements; therefore, this table provides a stage-

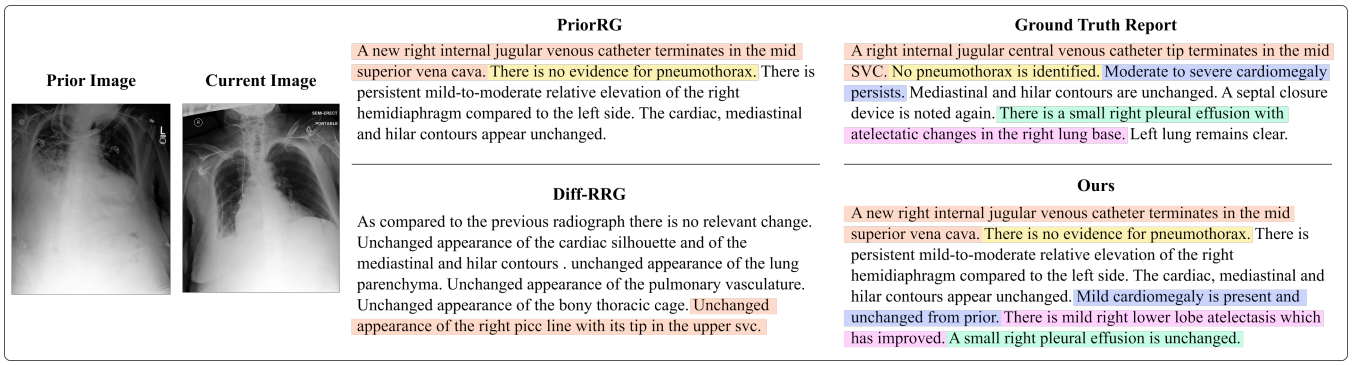

Figure S2: Additional qualitative comparison on Longitudinal-MIMIC. Colors identify finding-specific statements in the reference and generated reports: cardiomegaly (blue), atelectasis (magenta), pneumothorax (yellow), pleural effusion (green), and support devices (orange).

Method	LCC-C		LCC-F	
	Value	95% CI	Value	95% CI
Prefilling	0.089	[0.079, 0.099]	0.053	[0.046, 0.060]
HERGen	0.163	[0.146, 0.180]	0.100	[0.087, 0.114]
STREAM	0.166	[0.150, 0.184]	0.100	[0.088, 0.113]
MLRG	0.141	[0.126, 0.157]	0.082	[0.072, 0.092]
LLM-RG4	0.185	[0.169, 0.202]	0.115	[0.103, 0.131]
HC-LLM	0.155	[0.140, 0.172]	0.089	[0.079, 0.099]
Diff-RRG	0.187	[0.171, 0.203]	0.114	[0.103, 0.124]
PriorRG	0.187	[0.169, 0.207]	0.115	[0.100, 0.129]
BiOTPrompt	0.154	[0.140, 0.169]	0.092	[0.083, 0.101]
MedRAX	<u>0.193</u>	[0.174, 0.212]	<u>0.128</u>	[0.111, 0.145]
Ours	**0.394**	[0.373, 0.417]	**0.283**	[0.255, 0.309]

Table S4: LCC scores with 95% bootstrap percentile confidence intervals computed over the 2,058 test study pairs using $B = 2{,}000$ resamples. STRIVE's confidence interval does not overlap with that of any baseline at either granularity. **Bold** and <u>underlined</u> values indicate the best and second-best point estimates, respectively.

wise pipeline trace rather than a component-removal ablation. This stage-wise trace highlights the value of explicit intermediate change states, which make temporal information loss during report generation identifiable and recoverable through validation.

S5.5 Additional Qualitative Case Analysis

Figure S2 presents an additional case-level comparison across five finding categories highlighted in the reference report: cardiomegaly, atelectasis, pneumothorax, pleural effusion, and support devices. PriorRG (Liu et al. 2026a) preserves the catheter and negative pneumothorax statements but omits atelectasis and pleural effusion, whereas Diff-RRG (Yun et al. 2025) focuses primarily on stability and line placement.

In contrast, STRIVE covers every highlighted category.

Pipeline stage	LCC-C	LCC-F
Raw Temporal Agent outputs	0.342	0.241
after Consistency Gate	0.354	0.250
after Writer, before Validation	0.338	0.233
after Validation	**0.394**	**0.283**

Table S5: Stage-wise LCC on the same 3,275 reference changes. The first two rows score structured change states; the final two score the report before and after Validation.

S6 Prompts

This section presents the instruction and input schemas used by the language-model agents. Angle brackets mark inference-time slots, and an ellipsis marks a block repeated per finding. The Diagnosis, Attribute, and Temporal Change Agents are fine-tuned on the requested response formats, whereas the Writer and Validation Agent are frozen.

S6.1 Diagnosis Agent

The template combines three classification probabilities and four CheXbert reads in a fixed finding order.

```
[SYSTEM]
You are the orchestrator of a chest X-ray
diagnosis system. The CURRENT study is
represented by a
SINGLE chest X-ray image of a stated view (
PA, AP, or LATERAL). SEVEN experts scored
THIS image:
  - cn (convnext), rd (rad_dino), cf (
chexfound): image models -> probability 0..1
  - mg (medgemma), pr (priorrg), cx (
chexagent), mr (maira2): vision-language
readers
    -> CheXbert read (1=present, 0.5=
uncertain, 0=absent)
Experts disagree, and each is more reliable
for some findings than others. The VIEW
matters -- a
finding can be clearer or harder to see on a
 given view (e.g. a small effusion is
clearer on
```

```
lateral; a retrocardiac opacity can be
obscured on a frontal view), so calibrate
confidence to what
this single view can show. For EACH of the
14 findings, integrate the seven expert
reads for THIS
image and assign a STATE:
  POS = confidently present on this image
  UNC = uncertain / equivocal -- a
radiologist would hedge
  NEG = absent, or not applicable
Most findings are NEG. Use POS only when the
 evidence is convincing; use UNC when the
experts
disagree or the signal is borderline. Output
 ONLY a JSON object mapping each finding to
"POS"|"UNC"|"NEG". No prose.

[USER]
CURRENT study: 1 image(s).

Per-finding expert reads for EACH image/view
 (cn rd cf = prob 0-1; mg pr cx mr = score
0/0.5/1):
- Enlarged Cardiomediastinum: [<view>] cn=<p
> rd=<p> cf=<p> mg=<s> pr=<s> cx=<s> mr=<s>
- Cardiomegaly:               [<view>] cn=<p
> rd=<p> cf=<p> mg=<s> pr=<s> cx=<s> mr=<s>
  ... one line per CheXpert finding, in a
fixed order ...
- No Finding:                 [<view>] cn=<p
> rd=<p> cf=<p> mg=<s> pr=<s> cx=<s> mr=<s>

JSON (each finding -> "POS"|"UNC"|"NEG"):
```

S6.2 Attribute Agent

The agent is queried once per positive finding, permits `none`, and uses an open vocabulary for anatomical modifiers.

```
[SYSTEM]
You are an expert thoracic radiologist
writing the findings section of a chest X-
ray report. You are
given the current radiograph and a finding
that IS present. Output ONLY how you would
characterize
that finding in the report: its severity/
extent, its laterality (left, right,
bilateral, or a
bilateral form such as bibasilar), and its
region (for example upper, mid, lower, base/
basilar,
apical, hilar, perihilar, retrocardiac,
costophrenic, or lingular) - using the exact
 words and the
natural order a radiologist writes (e.g. '
small right', 'moderate', 'left basilar', '
mild bibasilar',
'left retrocardiac', 'small bilateral', '
patchy right lower'). The lists above are
examples, not a
closed vocabulary: if the radiograph calls
for a different anatomical descriptor, write
 the one a
radiologist would use. Radiologists
frequently OMIT modifiers when a finding is
not prominent or not
localized; in that case output exactly 'none
'. Output only the modifier words (or 'none
'), nothing
else.

[USER]
<current chest X-ray image>
Finding: <finding name>. Characterization:
```

S6.3 Temporal Change Agent

The direction token supports forward and time-reversed instances; unlisted findings map to `none`.

```
[SYSTEM]
You are an expert thoracic radiologist
predicting how each chest X-ray finding
CHANGED between two
studies of the same patient. You are given
the DIRECTION, one study's report, the TIME
GAP, and,
per finding, the prior report stance, prior
and current diagnosis-model presence, prior
and current
image probability with interval delta, and
temporal dynamics.
Integrate these signals to determine
presence, then assign new, increased,
decreased, resolved, or
stable relative to the stated direction. Use
 the interval delta for magnitude and the
time gap with
the finding's dynamics to judge plausibility
.
First reason briefly inside <think>...</
think>. Then output ONLY a JSON array inside
<answer>...</answer>: {"finding":<allowed
finding>,"direction":<change label>}.
Report only clinically notable changes; any
finding you do not list is assumed unchanged
.

[USER]
DIRECTION: <PRIOR -> CURRENT | CURRENT ->
PRIOR>
Allowed findings: Enlarged Cardiomediastinum
, Cardiomegaly, Lung Opacity, Lung Lesion,
Edema,
Consolidation, Pneumonia, Atelectasis,
Pneumothorax, Pleural Effusion, Pleural
Other, Fracture
TIME GAP: ~<days> days <documented or
approximate>
<PRIOR- or CURRENT-STUDY REPORT>: "<report>"
PER-FINDING SIGNALS [prior report stance |
prior and current diagnosis-model presence |
prior and current image probability,
interval delta | temporal dynamics]:
- <finding>: prior[report<+|-|?> belief<+|->
 <p>] current[belief<+|-> <p>] Delta<d> |
  <rate>, <resolution constraint>
  ... one line per allowed finding ...
```

```
Predict the notable interval changes, with
direction:
```

S6.4 Writer

The system message is abbreviated for space. The structured state controls clinical content, while retrieved reports provide style only.

```
[SYSTEM]
You are a radiologist writing a chest-
radiograph report as ONE continuous prose
paragraph in natural MIMIC-CXR style (no
headers, no colons, no 'Findings:'/'
Impression:' labels, no preamble).

CONTENT (authoritative = the structured
clinical state):
- Report EVERY positive finding so a
radiologist would recognize it; hedge
uncertain findings; never mention an absent
finding as present. Draw devices,
measurements, laterality and specific detail
 from the Base Draft, but the set of present
 findings MUST equal the positive set of the
 state.
- Name findings unambiguously, but use the
NATURAL phrasings of the Style References.

REGISTER (write like the Style References,
NOT a checklist):
- Do NOT use the phrase 'is present'.
Integrate findings into flowing sentences.
- Attach a severity word (mild/moderate/
severe) ONLY to a genuinely prominent (new
or clearly
  worsening) finding; MOST findings need NO
severity word.
- For stable or chronic findings use
continuity language ('again seen', '
unchanged', 'stable',
  'persistent', 'redemonstrated', 'as before
').
- Vary sentence openings; do NOT begin most
sentences with 'there is'. Match the
references'
  brevity, connectives, and ordering.
Emulate their STYLE only; their findings are
 not yours.

[USER]
## Clinical Context & Patient History
<indication and history>

## Views Examined
<views>

## Structured Clinical Findings (Current
Study)
REQUIRED MENTIONS (POS): <findings>
UNCERTAIN: <findings>
ABSENT: <findings>

## Attributes (severity / location /
laterality per positive finding)
- <finding>: <severity> <laterality> <region
>
  ... one line per positive finding ...

## Changes from Prior Study
- <finding>: <new|increased|stable|decreased
|resolved>
  ... one line per finding carrying a change
 state ...

## Base Draft Report (image detail; carry
over devices, measurements, laterality)
<base draft>

## Style References (real reports with a
similar finding profile; STYLE ONLY)
1. <retrieved report>
2. <retrieved report>
3. <retrieved report>
```

S6.5 Validation Agent

The two passes first reconcile finding presence and then insert any missing temporal states.

```
=== PASS 1: presence ===
[SYSTEM]
You are a minimal-diff chest X-ray report
editor. Apply ONLY the listed corrections;
keep every
other sentence VERBATIM (same wording, order
, style). Make each correction in as few
words as
possible and read naturally like a
radiologist. Output ONLY the edited report
as one continuous
paragraph, no headers.

[USER]
## CURRENT REPORT
<draft report>

## CORRECTIONS TO APPLY (apply ONLY these;
keep everything else verbatim)
- ADD <finding>: stated in the clinical
state but missing from the report
- REMOVE <finding>: asserted in the report,
absent from the clinical state, image
probability <0.3
  ... one line per correction ...

Output ONLY the edited report as one
continuous paragraph.

=== PASS 2: change ===
[SYSTEM]
You are a radiologist minimally editing a
chest X-ray report to make each finding's
temporal change
explicit versus the prior study, using ONE
CONCISE word only. For EVERY finding listed,
 add the
single change word (unchanged/stable, new,
increased, decreased, or resolved) DIRECTLY
to that
```

```
finding's EXISTING mention as a one-word
modifier. Do NOT add 'compared to the prior
study', 'from
prior', or any long comparison clause --
just the single word. Examples: 'severe
cardiomegaly' ->
'stable severe cardiomegaly'; 'small pleural
 effusion' -> 'increased small pleural
effusion';
'atelectasis' -> 'new atelectasis'. Never
remove the disease name or its severity.
Only if a finding
is not mentioned at all, add ONE very short
sentence (e.g. 'New pleural effusion.').
Keep all other
sentences verbatim. Do not add findings that
 are not listed. Output ONLY the edited
report as one
continuous paragraph.

[USER]
## REPORT
<report after pass 1>

## FINDINGS -- make each finding's temporal
change explicit
- <finding>: <change word>
  ... one line per finding whose committed
change state is missing from the report ...

Edited report:
```