ECG-RePAIR: Agentic Test-Time Repair with Evidence Grounding for ECG Interpretation

Anonymous authors
Paper under double-blind review

Abstract

Electrocardiography (ECG) is an essential diagnostic tool, where reliable interpretation requires clinical findings to be grounded in the underlying waveform. Recent multimodal large language models (MLLMs) have advanced automated ECG interpretation by integrating signal and image information, improving report completeness and diagnostic reasoning. However, these models still largely rely on one-pass generation, allowing uncertain or weakly grounded findings to persist in the final interpretation. In routine clinical practice, such findings are often revisited on the ECG waveform and reassessed against diagnostic criteria before an interpretation is finalized. Inspired by this practice, we propose ECG-RePAIR, an agentic test-time repair paradigm that organizes post-generation interpretation as a structured *evidence-to-action* process. ECG-RePAIR uses criterion-guided retrieval-augmented generation to determine what evidence should be examined, ECG tools to verify claims against the current waveform, and a repair advantage memory to determine whether the verified evidence warrants modifying the report. Extensive experiments show that ECG-RePAIR consistently improves diagnostic accuracy and report quality, while ablation studies and clinical expert evaluation further support its effectiveness.

1 Introduction

Electrocardiography (ECG) is central to the diagnosis and monitoring of cardiac disease, where reliable interpretation depends on connecting clinical findings to the underlying waveform (Wang et al., 2026; Hong et al., 2020). In practice, clinicians examine waveform features across relevant leads and assess them against established diagnostic criteria before reaching a conclusion. Automating this evidence-driven process has therefore become an important goal in ECG analysis. Recent multimodal large language models (MLLMs) have advanced this goal by integrating raw signals with rendered ECG images to produce more comprehensive clinical reports (Lan et al., 2025; Liu et al., 2024). Building on these advances, protocol-guided training further improves ECG interpretation by supervising structured diagnostic procedures and encouraging evidence-aware reasoning (Jin et al., 2026).

However, clinical interpretation does not end with the initial report. In routine practice, uncertain or potentially inconsistent findings are often revisited on the waveform before the interpretation is finalized. Clinicians typically return to the relevant leads, re-examine the underlying measurements, and reassess the finding against established diagnostic criteria before deciding whether an earlier conclusion should be retained or revised (Mason et al., 2007a; Rafie et al., 2021). This iterative review helps ensure that the final interpretation remains consistent with the observed ECG evidence. Current ECG MLLMs, by contrast, remain primarily optimized for one-pass generation and provide limited support for such post-generation re-evaluation. As a result, claims that appear plausible in language but are insufficiently grounded in the current ECG may persist in the final report.

Figure 1 illustrates this limitation with two examples using ECG-R1 (Jin et al., 2026) as the base model. In the first, an abnormal finding is interpreted as normal, leading to a missed diagnosis that becomes apparent only after the relevant waveform measurement is examined. In the second, an inappropriate lead-based interpretation produces the wrong conclusion,

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.

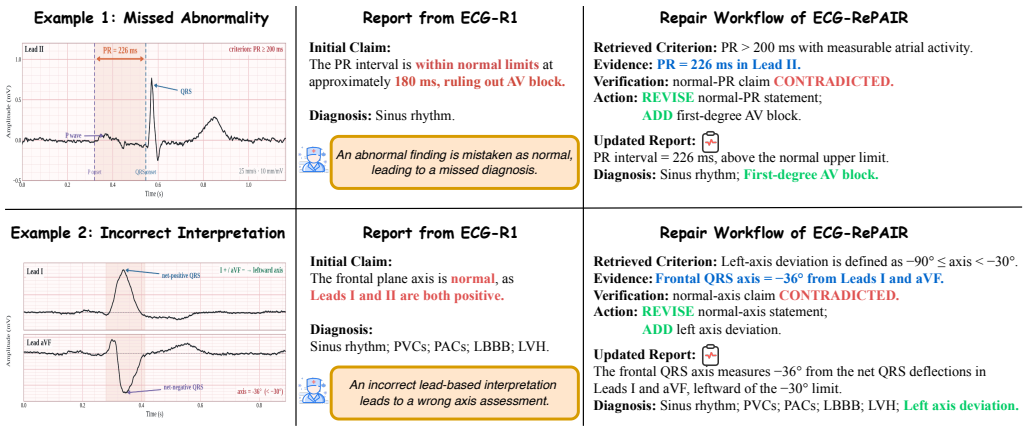


Figure 1: **Representative ECG report failures and localized repair.** Two examples illustrate errors from ECG-R1 (Jin et al., 2026) and how ECG-RePAIR resolves them through evidence-grounded local repair. See Appendix C for the complete repair workflow.

whereas criterion-guided measurement supports a different assessment. Despite arising from different interpretation errors, both cases expose the same underlying problem: a generated claim can appear clinically plausible without being explicitly checked against the waveform evidence required to support it. More importantly, detecting such a mismatch is only the first step. Reliable post-generation review must also determine what evidence should be examined and whether the verified finding warrants modifying the report.

To address these requirements, we introduce **ECG-RePAIR**, an agentic test-time repair framework inspired by routine clinical re-evaluation. It organizes post-generation ECG interpretation as a structured *evidence-to-action* process. ECG-RePAIR first performs *ECG-Aware Planning* to identify claims that warrant further inspection and uses criterion-guided retrieval-augmented generation (RAG) to determine what clinical evidence is required for each claim. It then conducts *Evidence-Grounded Verification*, where lead-aware ECG tools acquire the corresponding measurements from the current waveform and assess whether each claim is supported by the observed evidence. Finally, *Memory-Conditioned Repair* queries a repair advantage memory built from prior state–action outcomes to estimate whether an evidence-consistent report update is likely to improve the current interpretation.

This design combines current-waveform verification with experience-driven action selection: the observed ECG determines what is clinically supported, while accumulated repair experience helps adapt the current action decision. As a result, ECG-RePAIR grounds post-generation review in the present waveform while leveraging prior outcomes to make more effective repair decisions. Our contributions are summarized as follows:

- We propose ECG-RePAIR, an agentic test-time repair paradigm for ECG interpretation inspired by routine clinical re-evaluation. Rather than treating the initial report as the endpoint, ECG-RePAIR formulates post-generation interpretation as an *evidence-to-action* process in which uncertain findings can be re-examined before the final report is accepted.
- We realize this workflow through three ECG-specific mechanisms: criterion-guided RAG translates each claim into the evidence required for verification, lead-aware ECG tools obtain that evidence from the current waveform, and a repair advantage memory reuses prior state–action outcomes to adapt the current repair decision.
- Extensive evaluations show that ECG-RePAIR consistently improves diagnostic accuracy and report quality, while ablation studies and clinical expert evaluation further support the effectiveness of the proposed repair framework.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 presents ECG-RePAIR and its design. Section 4 reports the experimental setup and results, including comparative and ablation studies. Finally, Section 5 concludes the paper.

2 Related Work

ECG MLLMs. Multimodal large language models have rapidly expanded from general vision–language understanding (Hurst et al., 2024; Li et al., 2024) to medical reasoning and report generation (Li et al., 2023; Chen et al., 2024; Sellergren et al., 2025). ECG-specific MLLMs further adapt these capabilities to cardiac interpretation. ECG-Chat (Zhao et al., 2025), PULSE (Liu et al., 2024), and anyECG-Chat (Li et al., 2025a) support ECG diagnosis, reporting, and broader language interaction, while GEM (Lan et al., 2025) and ECG-R1 (Jin et al., 2026) further improve grounded and reliable report generation. Despite this progress, these methods still center on the initial interpretation, where unreliable conclusions may persist. Our work further advances ECG interpretation beyond generation through a grounded post-generation paradigm aligned with clinical re-evaluation.

Medical agentic reasoning. Medical MLLMs are increasingly moving beyond direct prediction toward structured reasoning and agentic diagnostic workflows. MedVLM-R1 (Pan et al., 2025), QoQ-Med (Dai et al., 2025), and Chiron (Sun et al., 2025) explore more structured medical reasoning and decision processes, while recent work on physiological waveforms highlights the potential of agentic reasoning for evidence-grounded ECG interpretation (Wang et al., 2026). Despite this progress, many existing medical agentic systems still simplify the iterative evidence review common in clinical practice. ECG-RePAIR further explores this direction for ECG interpretation through a clinically grounded agentic paradigm.

Verifiable ECG interpretation. Clinical ECG interpretation is inherently evidence-driven, relying on explicit measurements, lead-specific findings, and standardized diagnostic criteria (Kligfield et al., 2007; Mason et al., 2007b; Surawicz et al., 2009; Rautaharju et al., 2009). Existing signal-processing tools make such evidence directly measurable and reproducible (Makowski et al., 2021; Goldberger et al., 2000). Recent work highlights the importance of verifiable waveform reasoning for foundation models grounded in measurable physiological evidence (Wang et al., 2026). Motivated by this important direction, we explore a concrete realization of verifiable reasoning for comprehensive ECG interpretation.

3 Method

3.1 Overall Workflow

Clinical ECG interpretation rarely ends with an initial impression (Mason et al., 2007a; Rafie et al., 2021). When a finding appears uncertain or inconsistent, clinicians revisit the relevant waveform, compare the observed evidence with diagnostic criteria, and determine whether the interpretation should be retained or revised. Inspired by this clinical practice, we formulate ECG-RePAIR as an agentic test-time repair framework that moves ECG interpretation beyond initial generation through a structured *evidence-to-action* process.

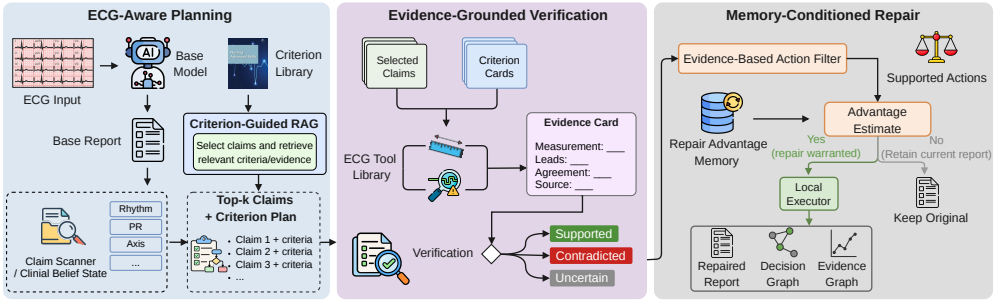


Figure 2: **Overview of ECG-RePAIR.** The framework proceeds through three stages: *ECG-Aware Planning* structures and prioritizes claims for inspection, *Evidence-Grounded Verification* evaluates them against waveform evidence, and *Memory-Conditioned Repair* uses prior repair experience to determine whether a localized report update is warranted.

As shown in Figure 2, ECG-RePAIR starts from an initial report produced by a frozen ECG model and maintains a clinical belief state over its findings. The repair process consists of three stages. (1) *ECG-Aware Planning* identifies claims that warrant further inspection and uses criterion-guided RAG to determine the evidence required for each claim. (2) *Evidence-Grounded Verification* acquires the corresponding waveform evidence with ECG tools and assesses whether the claim is supported by the current ECG. (3) *Memory-Conditioned Repair* queries a repair advantage memory built from prior state–action outcomes to determine whether an evidence-consistent report update is likely to improve the interpretation. The following subsections detail the design and implementation of each stage.

3.2 ECG-Aware Planning

ECG-RePAIR structures the initial interpretation into claim-level targets, prioritizes findings that warrant further inspection, and plans the evidence needed to verify them.

Claim-level state construction. Given an ECG input x with 12-lead waveform w, a base ECG model F_θ produces the initial report $y^{(0)} = F_\theta(x)$. A Claim Scanner decomposes the free-text report into addressable clinical findings, covering major ECG elements such as rhythm, conduction intervals, axis, voltage, and ST–T changes. Each finding is linked to the report span that expresses it and assigned a report status indicating whether it is currently asserted, denied, absent, or internally inconsistent. This step captures what the report says without judging whether the finding is physiologically correct.

The extracted claims are organized into a clinical belief state, which serves as the agent's structured working representation of the current interpretation. It summarizes the report context, the status of each claim, evidence acquired so far, unresolved findings, and the remaining query and edit budgets. At each decision step t, the Planner uses this evolving state to identify which claim would benefit most from additional waveform inspection. Rather than checking every possible ECG finding, it prioritizes a bounded set of candidate claims and passes the top-ranked ones within the query budget to evidence planning.

Criterion-guided planning. After the Planner selects a claim c_t for further inspection at decision step t, the next question is how that claim should be reliably verified. A high-level instruction such as "check this claim" is not yet actionable: verification requires knowing which leads to inspect, what quantity to measure, and which diagnostic criterion to apply. To provide this procedural guidance, ECG-RePAIR maintains a Clinical Criterion Library $\mathcal{K}$ grounded in standardized ECG interpretation recommendations (Kligfield et al., 2007; Mason et al., 2007b; Surawicz et al., 2009; Rautaharju et al., 2009).

Given the current clinical belief state, the Planner ranks candidate claims and selects the top-k claims within the query budget for further inspection. For each selected claim c_t, criterion-guided retrieval-augmented generation (RAG) retrieves the most applicable clinical rule from $\mathcal{K}$ and translates it into a structured *criterion-guided evidence plan*

$$p_t = \mathrm{Refine}(c_t, \mathcal{R}(\mathcal{K}, c_t)), \tag{1}$$

where $\mathcal{R}$ denotes the criterion retriever over $\mathcal{K}$, and Refine instantiates the retrieved rule for the selected claim by filling in the concrete lead set, target measurement, diagnostic threshold, and applicable prerequisites or exclusions. The resulting p_t therefore serves as an executable evidence plan for waveform inspection. For example, verifying a PR-interval claim may require measuring the PR interval in lead II, comparing it with the corresponding clinical boundary, and withholding judgment when atrial activity cannot be reliably assessed. The left part of Figure 3 illustrates this planning process with a concrete example.

The resulting plan turns a selected claim into an executable waveform inspection procedure. Importantly, retrieval specifies *what evidence should be obtained* without providing evidence about the current ECG itself. This retrieve–plan design follows broader retrieval-augmented and tool-oriented reasoning paradigms (Lewis et al., 2020; Asai et al., 2024; Yao et al., 2023; Schick et al., 2023), while restricting retrieval to ECG-specific criteria and measurement requirements. The finalized plan p_t is then passed to Evidence-Grounded Verification, where the required waveform evidence is directly acquired from the current ECG.

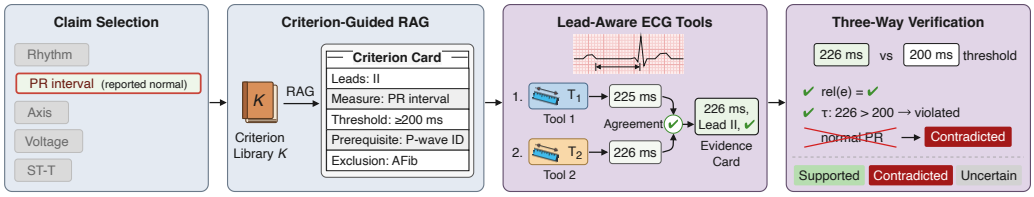


Figure 3: **Example of planning and verification.** For a PR-interval claim, ECG-RePAIR retrieves the relevant criterion, measures the required waveform evidence, and determines whether the claim is supported, contradicted, or uncertain.

3.3 Evidence-Grounded Verification

Given the criterion-guided evidence plan p_t, ECG-RePAIR executes the specified measurement on the current ECG and determines whether the resulting waveform evidence supports the selected claim. While the planning stage defines *what evidence is required*, verification grounds the assessment in observations obtained directly from the current waveform.

Lead-aware waveform measurement. Given the evidence plan p_t, ECG-RePAIR routes the specified leads and measurement target to the corresponding ECG tools. The tool library covers major ECG measurements, including rhythm, conduction intervals, frontal axis, voltage, and selected waveform morphology. To improve robustness, we use a multi-tool agreement design in which the same quantity is independently estimated by multiple signal-processing backends and retained only when their outputs are sufficiently consistent.

In our implementation, we use NeuroKit2 (Makowski et al., 2021) and the WFDB (Xie et al., 2023) delineator as two independent measurement backends. Both provide established ECG preprocessing and waveform delineation functions, with overlapping support for the measurements required by our criterion library. Using two independently implemented toolchains allows ECG-RePAIR to cross-check each target quantity rather than relying on a single measurement source. Let T_1 and T_2 denote the corresponding NeuroKit2- and WFDB-based measurement functions. Given waveform w and evidence plan p_t, their outputs are aggregated by $\mathcal{F}$ only when they agree within a quantity-specific tolerance:

$$e_t = \begin{cases} \mathcal{F}(T_1(w;p_t), T_2(w;p_t)), & \text{if } |T_1(w;p_t) - T_2(w;p_t)| \le \epsilon_\mu, \\ \perp, & \text{otherwise,} \end{cases} \tag{2}$$

where μ denotes the measurement type specified by p_t, ϵ_μ is its agreement tolerance in the corresponding measurement unit, and $\perp$ indicates that the measurement is unavailable for verification. Thus, disagreement between the two leads to withholding the observation.

Each accepted observation is represented as an Evidence Card e_t containing the measured value, unit, relevant leads, detector agreement, and provenance. The right part of Figure 3 illustrates this measurement and evidence construction process.

Three-way claim verification. The Clinical Verifier next compares the observed evidence e_t with the criterion specified by p_t. Let $\mathrm{rel}(e_t)$ indicate that the measurement is reliable, meaning that the backends agree, the required prerequisites are satisfied, and signal quality is sufficient. Let $\tau(e_t, p_t)$ indicate whether the observed value satisfies the corresponding clinical criterion. The claim status is then

$$z_t = \begin{cases} \textsc{Supported}, & \mathrm{rel}(e_t) \wedge \tau(e_t, p_t), \\ \textsc{Contradicted}, & \mathrm{rel}(e_t) \wedge \neg\tau(e_t, p_t), \\ \textsc{Uncertain}, & \text{otherwise.} \end{cases} \tag{3}$$

Poor signal quality, detector disagreement, missing prerequisites, or ambiguous near-boundary measurements therefore remain Uncertain rather than being forced into a binary judgment. The resulting evidence e_t and verification status z_t are written back to the clinical state for subsequent decision-making. In the evaluated configuration, at most six ranked claims are inspected, and no report content is modified during verification.

3.4 Memory-Conditioned Repair

Evidence-Grounded Verification determines what the current ECG supports, but not whether the report should change. ECG-RePAIR therefore formulates repair as an evidence-constrained decision problem: waveform evidence defines the admissible actions, while prior repair experience helps determine which action is worth taking.

Evidence-based action filter. For a verified claim c_t, its verification status z_t and current report status σ_t are passed through an evidence-based action filter to obtain the admissible action set $\mathcal{A}_t^E$. A supported finding missing from the report may admit Add, whereas reliable contradictory evidence may admit Revise. If the evidence is uncertain or already agrees with the current report, the default action is Keep. The filter therefore produces clinically evidence-consistent candidate actions without forcing a report update.

Repair advantage memory. ECG-RePAIR maintains a repair advantage memory $\mathcal{M}_{<t}$ built from prior repair episodes available before decision step t. Each prior episode records the clinical state, a candidate repair action, and the observed effect of that action on report utility. For a previous decision $i < t$, let y_i denote the report before action a_i and $y_i^{a_i}$ the locally updated report. Its observed advantage over retaining the original report is

$$\Delta_i = J(y_i^{a_i}) - J(y_i), \tag{4}$$

where $J(\cdot)$ is the scalar report-utility function defined by the evaluation protocol. Positive Δ_i indicates that the action improved report utility relative to Keep.

Each memory entry (s_i, a_i, Δ_i) therefore summarizes a prior state–action–advantage experience. The state descriptor s_i captures the information available at decision time, including the claim type, diagnostic context, verification status, measurement reliability, and distance from the relevant clinical boundary. Thus, $\mathcal{M}_{<t} = (s_i, a_i, \Delta_i)_{i<t}$ records how candidate repair actions performed under different verified clinical states without storing historical report text. At the current decision step t, the verified state s_t queries $\mathcal{M}_{<t}$ for relevant prior episodes. A memory-conditioned estimator then uses these matched state–action–advantage experiences to estimate the expected advantage of each admissible candidate action. Figure 4 illustrates this memory construction and query process.

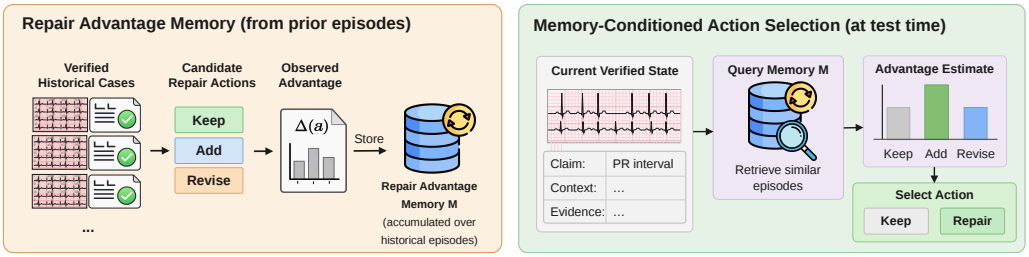


Figure 4: **Repair advantage memory and test-time action selection.** Prior verified repair episodes form state–action–advantage tuples in memory. At test time, the current verified state queries this memory to estimate action advantages and guide repair decisions.

Test-time action selection. For each $a \in \mathcal{A}_t^E$, the memory-conditioned estimator predicts its advantage relative to Keep as $\widehat{A}_t(a) = \widehat{\mathbb{E}}[\Delta \mid s_t, a; \mathcal{M}_{<t}]$. ECG-RePAIR selects

$$a_t^* = \arg\max_{a \in \mathcal{A}_t^E} \widehat{A}_t(a), \quad \text{execute } a_t^* \text{ only if } \widehat{A}_t(a_t^*) > m, \tag{5}$$

where m is a decision margin controlling how much estimated benefit is required before modifying the report. If no candidate action exceeds this margin, the current interpretation is retained. An accepted action is executed locally on the target claim while preserving unrelated findings. The final decision and its supporting waveform evidence are retained for traceability, without introducing additional diagnosis changes.

Table 1: Comparison of ECG interpretation performance (bold indicates the best result).

Model	Diag. Acc.	Credit	Compl.	Relev.	Lead	Ground.	Reason.	Fidelity
Proprietary MLLMs								
Gemini-3.8-Flash	61.34	58.76	5.52	4.77	5.33	67.11	68.57	69.52
GPT-5.6-Sol	64.75	59.36	4.56	5.35	5.63	72.56	71.38	72.13
Open-source MLLMs								
Qwen3-VL-8B-Instruct	6.26	4.88	1.10	0.45	0.41	18.82	18.73	18.90
Gemma-4-26B-A4B-it	9.74	8.32	1.64	0.56	0.53	35.24	31.45	35.66
MiniCPM-V-4.6	14.53	10.76	1.80	0.74	0.54	33.01	32.97	33.20
InternVL3-8B-Instruct	15.72	10.34	1.11	0.46	0.39	19.18	18.66	19.00
GLM-4.1V-9B-Base	16.36	13.52	2.17	1.06	1.19	42.00	41.56	41.77
MiMo-VL-7B-SFT	19.84	13.86	1.69	1.00	0.64	33.46	32.73	33.94
Medical MLLMs								
Chiron-o1-8B	11.13	7.79	0.95	0.67	0.22	20.24	18.88	19.37
HuatuoGPT-Vision-7B	12.75	9.36	3.43	1.12	0.67	49.44	49.72	49.72
MedVLM-R1	26.88	17.10	1.56	0.55	0.60	25.17	24.89	25.02
QoQ-Med-VL-7B	27.64	23.56	2.06	1.70	1.28	35.01	34.96	35.20
ECG-specialized MLLMs								
PULSE-7B	60.98	48.73	1.93	2.50	0.69	47.42	47.49	47.49
GEM	65.88	56.53	4.30	5.21	6.23	64.71	64.62	65.03
ECG-R1-8B-RL	76.24	68.65	4.62	5.78	5.82	71.41	71.37	72.16
+ECG-RePAIR (Ours)	**80.49**	**74.94**	**5.24**	**6.33**	**6.42**	**75.09**	**75.04**	**75.32**

4 Experiments

4.1 Performance Evaluations

Dataset. We evaluate on the ECG-Grounding test set derived from MIMIC-IV-ECG (Johnson et al., 2023), following the setting of ECG-R1 (Jin et al., 2026). The released test index contains 2,381 ECG records, each with a 12-lead waveform and a reference interpretation. We retain the 2,380 records with valid ECG-R1 judgments for all experiments.

Evaluation metrics. Following ECG-R1, we report eight standard evaluation metrics covering diagnostic correctness and report quality: Diagnosis Accuracy, Diagnosis Credit, Analysis Completeness, Analysis Relevance, Lead-Evidence Validity, ECG-Feature Grounding, Evidence-Based Reasoning, and Clinical Diagnostic Fidelity. These metrics are computed using the enumerated ECG-R1 rubric with `DeepSeek-v4-flash-0731` as the judge. All methods are scored with the same rubric and judge on the common 2,380-record test set.

Compared methods. We compare ECG-RePAIR with four groups of multimodal models. Proprietary MLLMs serve as general-purpose commercial baselines. Open-source MLLMs include Qwen3-VL, MiniCPM-V, InternVL3, GLM-4.1V, and MiMo-VL (Bai et al., 2025; Yao et al., 2024; Zhu et al., 2025; GLM-V Team et al., 2025; Li et al., 2025b). Medical MLLMs include Chiron-o1, HuatuoGPT-Vision, MedVLM-R1, and QoQ-Med (Sun et al., 2025; Chen et al., 2024; Pan et al., 2025; Dai et al., 2025). ECG-specialized models include PULSE, GEM, , and ECG-R1-8B-RL (Liu et al., 2024; Lan et al., 2025; Jin et al., 2026), where ECG-R1-8B-RL denotes the 8B-parameter reinforcement-learning variant of ECG-R1.

Implementation details. ECG-RePAIR [1] takes the released ECG-R1-8B-RL as the base model. The five-fold protocol is used specifically to construct and query the repair advantage memory without exposing a record to its own outcome. We deterministically partition the 2,380 records into five folds. For each held-out fold, the remaining four folds are used to construct the memory, fit the action estimator, and select the decision margin; the held-out fold is used only for test-time repair and evaluation. Rotating the held-out fold yields out-of-fold ECG-RePAIR results for all 2,380 records. Thus, every repair is conditioned only on experience from other records, while the final comparison with all baselines is still performed on the same complete test set. Additional details are provided in Appendix A.

[1] Our code is available at `https://anonymous.4open.science/r/ECG-RePAIR-D5C6/`.

Results and analysis. Table 1 summarizes the comparison results. ECG-RePAIR improves its frozen base across all eight metrics, increasing Diagnosis Accuracy from 76.24 to 80.49 and Diagnosis Credit from 68.65 to 74.94. The gains in Lead-Evidence Validity and ECG-Feature Grounding indicate more reliable use of lead-level waveform evidence, while improvements in Evidence-Based Reasoning and Clinical Diagnostic Fidelity reflect stronger evidence-supported reasoning and better alignment between the final interpretation and the underlying ECG. ECG-specialized models also generally outperform general-purpose and medical MLLMs, underscoring the value of ECG-specific modeling for this task.

4.2 Ablation Study

We evaluate the three key components of ECG-RePAIR in their execution order: criterion-guided planning, ECG tool-based verification, and memory-conditioned repair, with each component building upon the previous stage. All variants use the same evaluation setting. Specifically, `w/o RAG` removes criterion retrieval and instead uses fixed claim-specific verification rules. `w/o ECG Tools` retains the retrieved criterion but replaces waveform measurements from ECG tools with model-based verification by the base ECG model. `w/o Memory` retains criterion-guided planning and waveform verification, but removes memory-conditioned action selection and executes admissible repair actions in a fixed order.

Results and analysis. As shown in Table 2, removing any of the three components reduces performance across all evaluation metrics. Criterion-guided planning improves the quality of evidence selection, while ECG tool-based verification provides the largest gain by grounding decisions in direct waveform measurements. Building on these two stages, memory-conditioned repair further improves performance by selecting which evidence-consistent actions are worth executing. These results validate the effectiveness of the individual components and their complementary roles in the sequential evidence-to-action process.

Table 2: Ablation study of ECG-RePAIR.

	Components			Evaluation metrics							
Variant	RAG	Tools	Memory	Diag. Acc.	Credit	Compl.	Relev.	Lead	Ground.	Reason.	Fidelity
ECG-RePAIR	✓	✓	✓	**80.49**	**74.94**	**5.24**	**6.33**	**6.42**	**75.09**	**75.04**	**75.32**
w/o RAG	×	✓	✓	79.31	73.42	5.02	6.08	6.16	74.21	74.08	74.35
w/o ECG Tools	✓	×	✓	78.22	71.83	4.81	5.84	5.57	72.14	72.31	72.68
w/o Memory	✓	✓	×	79.76	73.88	5.16	6.18	6.31	74.62	74.48	74.71

4.3 Robustness Evaluation

We conduct two additional experiments to examine whether the effectiveness of ECG-RePAIR depends on the choice of base ECG model or evaluator. For the base-model study, we replace ECG-R1 with GEM while keeping the verification-and-repair framework and evaluation protocol unchanged. For the evaluator study, we introduce the more recent `Qwen3.8-Flash-Next-FP8` as an additional judge from a different model family. The generated ECG-R1 and ECG-RePAIR reports are kept fixed and rescored using the same evaluation rubric, with all other settings unchanged.

Results and analysis. As shown in Figure 5, applying ECG-RePAIR to GEM improves Diagnosis Accuracy from 65.88 to 67.85 and yields consistent gains across the remaining metrics, indicating that the repair framework can generalize beyond the ECG-R1 base model. The smaller gain relative to ECG-R1 further suggests that repair effectiveness may depend on the structure and addressability of the initial report. When the fixed reports are re-evaluated by `Qwen3.8-Flash-Next-FP8`, Diagnosis Accuracy increases from 65.52 to 69.70, with consistent improvements across all eight metrics. Although the absolute scores differ from those produced by the primary judge, the overall improvement trend remains unchanged. Together, these results show that ECG-RePAIR remains effective under changes to both the base ECG model and the evaluation model.

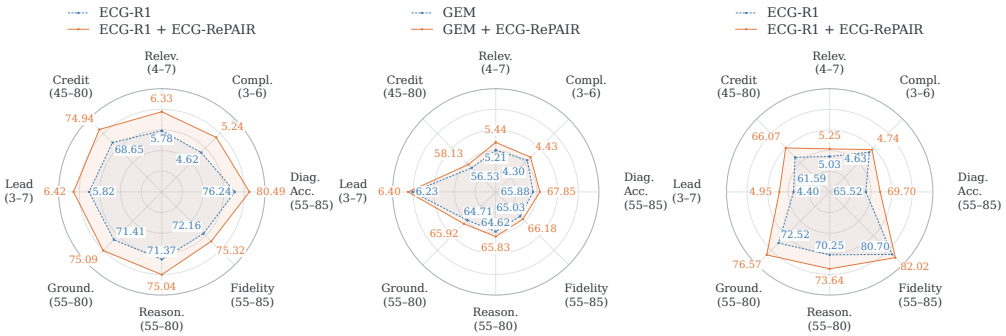


Figure 5: Robustness evaluation. From left to right: the main evaluation setting, GEM as the alternative base model, and a Qwen model as the alternative evaluator.

4.4 Clinical Expert Evaluation

We conduct an additional evaluation with two clinicians to assess the clinical quality of ECG-RePAIR's report modifications. Each clinician independently reviews the same 50 repaired cases and evaluates whether the modified claim is supported by the ECG, whether the repair improves the report, and whether it introduces a new error. Individual assessments are retained without resolving disagreements into a consensus label.

Results and analysis. As shown in Figure 6, Expert 1 and Expert 2 judge the modified claims to be supported in 82% and 86% of cases, respectively, and consider the repaired reports improved in 74% and 78% of cases. Exact agreement reaches 92% for claim support, 90% for clinical effect, and 92% for newly introduced errors. These results show that most sampled repairs are supported by the ECG and are judged clinically beneficial, while only a small number introduce potentially problematic changes.

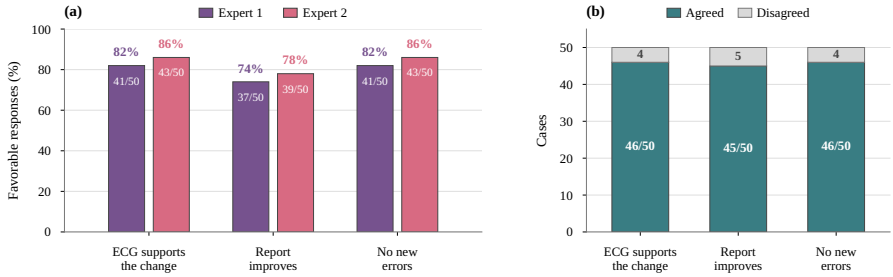


Figure 6: Clinical expert evaluation of 50 repaired reports. (a) Favorable-response rates for claim support, report improvement, and absence of new errors. (b) Agreement between the two experts across the original response categories.

5 Conclusion

We presented ECG-RePAIR, an agentic test-time repair framework that extends ECG interpretation beyond one-pass generation through a structured *evidence-to-action* process. By combining criterion-guided planning, waveform-grounded verification, and memory-conditioned repair, ECG-RePAIR reassesses uncertain findings and determines whether the initial report should be updated. Extensive evaluations show improvements in diagnostic accuracy and report quality, while ablation studies and clinical expert evaluation further support the effectiveness of the framework. A current limitation is that ECG-RePAIR operates only after the initial report is generated and leaves the underlying model unchanged; future work could incorporate the same evidence-to-action principle into training so that waveform verification and repair-aware reasoning become intrinsic model capabilities.

AI use statement

We used generative AI tools to assist with language polishing and to improve the clarity and readability of the manuscript. All AI-assisted text was reviewed and revised by the authors. Generative AI tools were not used to generate experimental results or replace human evaluation. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This study uses de-identified ECG data from existing research datasets and does not involve the collection of new patient data. The proposed framework is intended for research on automated ECG interpretation and is not evaluated for clinical deployment or autonomous medical decision-making. Clinical expert evaluation is used only to assess generated report modifications on the de-identified study cases. We acknowledge that errors in automated ECG interpretation may have clinical consequences, and further validation would be required before any real-world use.

Reproducibility statement

We provide detailed descriptions of the dataset, evaluation protocol, model settings, ECG tools, clinical criterion library, memory construction, and hyperparameters in the main paper and appendix. The implementation and supporting code are available in our anonymous repository: https://anonymous.4open.science/r/ECG-RePAIR-D5C6/.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. HuatuoGPT-Vision: Towards injecting medical visual knowledge into multimodal LLMs at scale. *arXiv preprint arXiv:2406.19280*, 2024.

Jacqueline Christianson. *An EKG Interpretation Primer*. Nurses International Team, 2019. URL https://open.umn.edu/opentextbooks/textbooks/an-ekg-interpretation-primer. CC BY 4.0.

Wenrui Dai, Pengrui Chen, Chirag Ekbote, and Paul Pu Liang. QoQ-Med: Building multimodal clinical foundation models with domain-aware GRPO training. *arXiv preprint arXiv:2506.00711*, 2025.

Kim Ernstmeyer and Elizabeth Christman, editors. *Nursing Advanced Skills*. Chippewa Valley Technical College / Open Resources for Nursing, 2023. URL https://www.ncbi.nlm.nih.gov/books/NBK594493/. Chapter 7: Interpret Basic ECG; CC BY 4.0.

GLM-V Team, Wenyi Hong, Wenmeng Yu, et al. GLM-4.1V-Thinking and GLM-4.5V: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Google DeepMind. Gemini 3.8 flash model card. https://deepmind.google/models/model-cards/gemini-3-8-flash/, September 2026. Released September 2, 2026; accessed September 25, 2026.

E. William Hancock, Barbara J. Deal, David M. Mirvis, et al. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part v: Electrocardiogram changes associated with cardiac chamber hypertrophy. *Journal of the American College of Cardiology*, 53(11):992–1002, 2009. doi: 10.1016/j.jacc.2008.12.015.

Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, 122:103801, 2020.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jiarui Jin, Haoyu Wang, Xingliang Wu, Xiaocheng Fang, Xiang Lan, Zihan Wang, Deyun Zhang, Bo Liu, Yingying Zhang, Xian Wu, Hongyan Li, and Shenda Hong. ECG-R1: Protocol-guided and modality-agnostic MLLM for reliable ECG interpretation. *arXiv preprint arXiv:2602.04279*, 2026.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, and Brian Gow. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Paul Kligfield, Leonard S. Gettes, James J. Bailey, et al. Recommendations for the standardization and interpretation of the electrocardiogram: Part i: The electrocardiogram and its technology. *Journal of the American College of Cardiology*, 49(10):1109–1127, 2007. doi: 10.1016/j.jacc.2007.01.024.

Xiang Lan, Feng Wu, Kai He, Qinghao Zhao, Shenda Hong, and Mengling Feng. GEM: Empowering MLLM for grounded ECG understanding with time series and images. *arXiv preprint arXiv:2503.06073*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K"uttler, Mike Lewis, Wen-tau Yih, Tim Rockt"aschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Haitao Li, Ziyu Li, Yiheng Mao, Zhixuan Liu, Zhaoyang Sun, and Zhen Huang. anyECG-chat: A generalist ECG-MLLM for flexible ECG input and multi-task understanding. *arXiv preprint arXiv:2506.00942*, 2025a.

J. Li, J. Chen, Y. Qu, S. Xu, Z. Lin, J. Zhu, B. Xu, W. Tan, P. Fu, J. Ju, Z. Luo, and J. Luan. Xiaomi MiMo-VL-Miloco technical report. *arXiv preprint arXiv:2512.17436*, 2025b.

Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. Teach multimodal LLMs to comprehend electrocardiographic images. *arXiv preprint arXiv:2410.19008*, 2024.

Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4):1689–1696, 2021.

Jay W. Mason, E. William Hancock, Leonard S. Gettes, et al. Recommendations for the standardization and interpretation of the electrocardiogram: Part ii: Electrocardiography diagnostic statement list. *Circulation*, 115(10):1325–1332, 2007a. doi: 10.1161/CIRCULATIONAHA.106.180201.

Jay W. Mason, E. William Hancock, Leonard S. Gettes, et al. Recommendations for the standardization and interpretation of the electrocardiogram: Part ii: Electrocardiography diagnostic statement list. *Circulation*, 115(10):1325–1332, 2007b. doi: 10.1161/CIRCULATIONAHA.106.180201.

OpenAI. GPT-5.6 system card. https://deploymentsafety.openai.com/gpt-5-6, 2026. Accessed September 2026.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.

Nikita Rafie, Anthony H. Kashou, and Peter A. Noseworthy. Ecg interpretation: Clinical relevance, challenges, and advances. *Hearts*, 2(4):505–513, 2021. doi: 10.3390/hearts2040039.

Pentti M. Rautaharju, Borys Surawicz, and Leonard S. Gettes. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part iv: The ST segment, T and U waves, and the QT interval. *Journal of the American College of Cardiology*, 53(11):982–991, 2009. doi: 10.1016/j.jacc.2008.12.014.

Timo Schick, Jane Dwivedi-Yu, Roberto Dess'i, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Hao Sun, Yifan Jiang, Wei Lou, Yiming Zhang, Wenrui Li, Lili Wang, Meng Liu, Lin Liu, and Xiang Wang. Chiron-o1: Igniting multimodal large language models towards generalizable medical reasoning via mentor–intern collaborative search. In *Advances in Neural Information Processing Systems*, 2025.

Borys Surawicz, Rory Childers, Barbara J. Deal, et al. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part iii: Intraventricular conduction disturbances. *Journal of the American College of Cardiology*, 53(11): 976–981, 2009. doi: 10.1016/j.jacc.2008.12.013.

Galen S. Wagner, Peter Macfarlane, Hein Wellens, et al. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part vi: Acute ischemia/infarction. *Journal of the American College of Cardiology*, 53(11):1003–1011, 2009. doi: 10.1016/j.jacc.2008.12.016.

Xiaoda Wang, Ching Chang, Defu Cao, Kaiqiao Han, Fang Sun, Yue Huang, Minxiao Wang, Chang Xu, Xiao Luo, Runze Yan, et al. Position: Beyond prediction: Toward verifiable physiological waveform reasoning with foundation models and agentic llms. 2026.

Chen Xie, Lucas McCullum, Alistair Johnson, Tom Pollard, Brian Gow, and Benjamin Moody. Waveform database software package (wfdb) for python. *PhysioNet*, January 2023. doi: 10.13026/9njx-6322. URL https://doi.org/10.13026/9njx-6322. Version 4.1.0.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. MiniCPM-V: A GPT-4V level MLLM on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. ECG-Chat: A large ECG-language model for cardiac disease diagnosis. *arXiv preprint arXiv:2408.08849*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Details of Experimental Setup

This section provides further details of the experimental setup, including the dataset, evaluation metrics and prompts, compared methods, evaluation protocol, ECG-RePAIR implementation, and clinical expert evaluation protocol.

A.1 Dataset

The ECG-Grounding test set is derived from MIMIC-IV-ECG (Johnson et al., 2023). MIMIC-IV-ECG contains approximately 800,000 diagnostic ECGs from nearly 160,000 patients, with each recording comprising a 10-second, 12-lead waveform sampled at 500 Hz. ECG-R1 (Jin et al., 2026) constructs an ECG-Grounding test index of 2,381 records from this resource, where each case pairs a 12-lead ECG waveform with a reference interpretation. Among these records, 2,380 have valid ECG-R1 judgments and are used in our primary experiments.

A.2 Evaluation Metrics

Following ECG-R1 (Jin et al., 2026), we evaluate generated ECG reports using eight rubric-based metrics covering diagnostic correctness and report quality: Diagnosis Accuracy, Diagnosis Credit, Analysis Completeness, Analysis Relevance, Lead-Evidence Validity, ECG-Feature Grounding, Evidence-Based Reasoning, and Clinical Diagnostic Fidelity.

Evaluation prompts. Each metric is scored by an LLM judge using the corresponding ECG-R1 rubric and metric-specific prompt. We use `DeepSeek-v4-flash-0731` as the primary judge model and apply the same evaluation configuration across all methods within each comparison. The Base ECG Interpretation Prompt, Grounded ECG Interpretation Evaluation Prompt, and Judge System Prompt are provided below.

Base ECG Interpretation Prompt

Interpret the provided ECG image, identify key features and abnormalities in each lead, and generate a clinical diagnosis that is supported by the observed evidence.

ECG Interpretation Evaluation Prompt

Your task: Evaluate the alignment and quality of a generated ECG interpretation by comparing it to a ground truth clinician's interpretation.

Evaluation Criteria:

1. DiagnosisAccuracy: Evaluates whether the generated diagnosis is correct, specific, and supported by ECG findings.

Scoring

+**2 per diagnosis:** Each correctly identified key diagnosis with supporting ECG features.

+**1 per diagnosis:** Each mostly correct diagnosis but lacking key supporting details.

+**0 per diagnosis:** Each incorrect or vague diagnosis not supported by ECG features.

2. AnalysisCompleteness: Checks if all key ECG components (rhythm, intervals, waveforms, and lead-specific findings) are discussed.

Scoring

+**1 per feature:** For each correctly addressed key ECG feature (e.g., rhythm, PR interval, QRS duration, ST segment, T wave morphology).

+**0 per missing feature:** For each key feature omitted or inaccurately described.

3. AnalysisRelevance: Assesses whether each provided explanation directly supports the diagnosis.

Scoring

+**2 per feature or per lead:** Each point that strongly supports the diagnosis with clear ECG evidence.

+**1 per feature or per lead:** Some points are relevant but not fully justified.

+**0:** Includes unrelated or misleading explanations.

4. LeadEvidenceValidity: Evaluates whether the lead-related statements are diagnostically necessary, correctly grounded, and free of unsupported lead-wise claims, rather than maximizing the number of mentioned leads.

Scoring

+**2 per key lead/region:** For each diagnosis-critical lead (or contiguous lead group / territory) correctly referenced with explicit and matching ECG evidence (e.g., ST elevation/depression, T-wave inversion, pathologic Q waves, bundle-branch morphology) that supports the stated diagnosis.

+**1 per key lead/region:** For each key lead/region mentioned with partially correct evidence, but missing important details (e.g., direction without magnitude/morphology) or showing minor inconsistencies.

+**0 per key lead/region:** Key lead/region is omitted or the described finding is incorrect / non-grounded and does not support the diagnosis.

5. GroundedECGUnderstanding: Determines if the interpretation references actual ECG features (e.g., QRS amplitude, PR interval) instead of generic terms.

Scoring (0–100)

100: ECG findings are comprehensively cited, linked to diagnoses, and cover all relevant ECG features.

80: ECG findings are explicitly cited and linked to diagnoses.

50: Some ECG references exist but are incomplete.

0: Lacks specific waveform references.

6. EvidenceBasedReasoning: Evaluates whether the diagnosis follows logical, evidence-supported steps.

Scoring (0–100)

100: Findings logically progress to diagnosis with thorough and clear justifications covering all necessary steps.

80: Findings logically progress to diagnosis with clear justifications.

50: Some reasoning exists but lacks complete step-by-step analysis.

0: Reasoning is unclear or not derived from ECG findings.

7. RealisticDiagnosticProcess: Assesses if the model mimics how a clinician interprets an ECG, considering all relevant factors.

Scoring (0–100)

100: The analysis follows a structured clinical approach and considers all relevant clinical factors.

80: The analysis follows a structured clinical approach.

50: Some clinical reasoning is present but incomplete.

0: The approach lacks structured clinical reasoning.

NOTE: Each score must be calculated based on strict criteria to ensure objective evaluation.

Generated ECG Interpretation:

{generated}

Ground Truth Clinician's Interpretation:

{groundtruth}

Response: [Please organize your output in a JSON format for each criterion, including a brief explanation for each aspect. Strictly follow this JSON format that records every scoring for each criterion:

{{'DiagnosisAccuracy':[{{'Score': , 'Explanation':}}, ...],

'AnalysisCompleteness':[{{'Score': , 'Explanation':}}, ...],

'AnalysisRelevance':[{{'Score': , 'Explanation':}}, ...],

'LeadEvidenceValidity':[{{'Score': , 'Explanation':}}, ...],

'GroundedECGUnderstanding':[{{'Score': , 'Explanation':}}, ...],

'EvidenceBasedReasoning':[{'Score': , 'Explanation':}}, ...],
'RealisticDiagnosticProcess':[{'Score': , 'Explanation':}}, ...]}}]

Judge System Prompt

You are grading with the rubric in the user message. Follow its counting rule exactly: DiagnosisAccuracy is scored '+2 per diagnosis', '+1 per diagnosis', '+0 per diagnosis', and AnalysisCompleteness, AnalysisRelevance and LeadEvidenceValidity are likewise scored per feature, per explanation and per lead. Emit one JSON array entry for each individual diagnosis, feature, explanation and lead you score – never a single pooled entry summarising several of them. Score every diagnosis the generated interpretation asserts, including ones the ground truth does not mention, which take +0 when they are unsupported. Keep each Explanation brief.

A.3 Compared Methods

We compare ECG-RePAIR against four groups of multimodal models:

- **General-purpose proprietary MLLMs:** Gemini-3.8-Flash (Google DeepMind, 2026) and GPT-5.6-Sol (OpenAI, 2026).
- **Open-source MLLMs:** Qwen3-VL-8B-Instruct (Bai et al., 2025), Gemma-4-26B-A4B-it, MiniCPM-V-4.6 (Yao et al., 2024), InternVL3-8B-Instruct (Zhu et al., 2025), GLM-4.1V-9B-Base (GLM-V Team et al., 2025), and MiMo-VL-7B-SFT (Li et al., 2025b).
- **Medical MLLMs:** Chiron-o1-8B (Sun et al., 2025), HuatuoGPT-Vision-7B (Chen et al., 2024), MedVLM-R1 (Pan et al., 2025), and QoQ-Med-VL-7B (Dai et al., 2025).
- **ECG-specialized MLLMs:** PULSE-7B (Liu et al., 2024), GEM (Lan et al., 2025), and ECG-R1-8B-RL (Jin et al., 2026), the 8B-parameter RL variant of ECG-R1.

Unless otherwise specified, all compared methods use their released checkpoints and default preprocessing and inference settings.

A.4 Evaluation Protocol

ECG-RePAIR takes the released ECG-R1-8B-RL reports as its initial interpretations and keeps the base model frozen throughout repair. We use a five-fold protocol to construct and query the repair advantage memory while preventing each record from using its own evaluation outcome. The 2,380 records are deterministically partitioned into five folds. For each held-out fold, the remaining four folds are used to construct the repair advantage memory, fit the action estimator, and select the decision margin, while the held-out fold is used only for test-time repair and evaluation. Rotating the held-out fold produces out-of-fold ECG-RePAIR results for the full evaluation cohort.

Each memory episode stores the verified state, candidate repair action, and its observed report-utility change relative to retaining the original report. Historical report text, reference interpretations, and judge outputs are excluded from the input features used for the current repair decision. The memory is constructed before evaluation and remains read-only during each held-out pass, so test records are not added back to memory during inference.

A.5 ECG-RePAIR Implementation Details

Base ECG model. ECG-RePAIR uses ECG-R1-8B-RL[2], the released 8-billion-parameter reinforcement-learning variant of ECG-R1 (Jin et al., 2026), to generate the initial ECG report. The base-model parameters remain fixed throughout the entire repair process.

[2] https://huggingface.co/PKUDigitalHealth/ECG-R1-8B-RL

Clinical criterion library. ECG-RePAIR maintains a structured criterion library that maps each ECG finding to the evidence required for verification. Each entry specifies the target measurement, relevant leads, clinical boundary, and applicable validity conditions. Given a selected claim, criterion-guided retrieval identifies the relevant clinical rule and instantiates it as an executable evidence plan.

The library is grounded in established ECG interpretation recommendations (Kligfield et al., 2007; Surawicz et al., 2009; Rautaharju et al., 2009; Hancock et al., 2009; Wagner et al., 2009), which provide the core measurement definitions, diagnostic boundaries, and applicable interpretation criteria. Openly available ECG educational resources are further used as supplementary explanatory context (Ernstmeyer and Christman, 2023; Christianson, 2019). Retrieved information is then consolidated into a structured Criterion Card that specifies the evidence needed for subsequent waveform verification. A concrete example is shown below.

Example of Criterion Retrieval and Card Instantiation

Retrieved passage. "First-degree AV block is defined as a prolonged PR interval more than 0.2 seconds. A single P wave precedes every QRS complex by a consistent length."

Instantiated Criterion Card

Claim: prolonged PR interval / first-degree AV block
Measurement: PR interval in milliseconds
Relevant lead: lead II, or the lead with the clearest measurable P-wave onset
Clinical boundary: PR interval greater than 200 ms
Prerequisites: identifiable P-wave onset and one atrial activation preceding each measured QRS complex
Exclusions: atrial fibrillation or flutter, poorly visible P waves, insufficient signal quality, or unresolved disagreement between ECG tools
Planner output: the lead to inspect, the measurement to obtain, the criterion to apply, and the validity conditions required for verification

ECG toolset. Waveform measurements are obtained from the current digital ECG using two measurement backends. Given a selected claim, its Criterion Card specifies the measurement to perform and the leads to inspect. For interval-related claims, the NeuroKit2 path cleans lead II, detects R peaks, and applies wavelet-based delineation to locate P- and T-wave landmarks. The WFDB/XQRS path independently detects R peaks in lead II and applies a morphology-based search for the corresponding waveform landmarks (Makowski et al., 2021; Goldberger et al., 2000). Shared measurement routines then compute the requested quantity from the relevant leads, including rate and rhythm, conduction intervals, frontal axis, voltage, and selected morphology and ST–T features. The two backend measurements are compared using a quantity-specific tolerance. When they agree, the result is recorded in an Evidence Card containing the value, unit, lead set, backend agreement, and provenance; otherwise, the evidence is treated as unavailable for verification.

Example of ECG Tool Execution

Criterion Card: measure PR interval in lead II; compare with the 200 ms boundary.

NeuroKit2 path: signal cleaning → R-peak detection → wavelet delineation → PR measurement.
WFDB/XQRS path: XQRS peak detection → morphology-based landmark search → PR measurement.

Evidence Card: PR value, ms, lead II, backend agreement, and waveform provenance.

Memory-based advantage estimation. ECG-RePAIR uses the repair advantage memory to estimate the expected benefit of each admissible repair action under the current verified state. Concretely, we fit a ridge-regression model on historical repair episodes, where the input is the state–action pair (s_i, a_i) and the regression target is its observed repair advantage Δ_i. The verified state is encoded using compact clinical information available at inference

time, including claim identity, diagnostic context, verification status, measurement reliability, tool agreement, and available measurement or boundary-margin information. At test time, the current verified state s_t is paired with each admissible action a, and the fitted regression model predicts the corresponding advantage $\widehat{A}_t(a)$. These predicted advantages are then used to rank candidate actions and determine whether a repair should be executed. Model fitting uses only repair episodes from the non-held-out folds.

Parameter settings. ECG-RePAIR uses a small set of parameters for claim inspection, memory-conditioned action selection, and ECG tool agreement. Table 3 summarizes both the fixed settings and the search spaces used during parameter selection.

Table 3: **Parameter settings of ECG-RePAIR.**

Setting	Value or search space
Maximum inspected claims per ECG	6
Admissible repair actions	KEEP, ADD, REVISE
Maximum executed repair actions	1
Advantage estimator	Ridge regression
State representation	Compact clinical features
Ridge coefficient α	$\{0.1, 1, 10, 100\}$
Decision margin m	$\{0, 1, 2, 5, 10\}$

The ridge coefficient α and decision margin m are selected separately for each held-out fold from the candidate sets above using only repair episodes from the remaining folds. For folds $\{0, 1, 2, 3, 4\}$, the selected values are $\alpha = \{100, 100, 100, 10, 10\}$ and $m = \{2, 2, 1, 0, 0\}$, respectively. The compact state representation is used in all five folds.

Table 4 lists the preset agreement tolerances used by the two ECG measurement backends for each supported quantity. For a measurement type μ, the two outputs $T_1^{(\mu)}$ and $T_2^{(\mu)}$ are considered consistent when

$$\left| T_1^{(\mu)} - T_2^{(\mu)} \right| \leq \epsilon_\mu, \tag{6}$$

where ϵ_μ denotes the corresponding tolerance in Table 4. Measurements that fail this agreement check are treated as unavailable for claim verification. These tolerances are fixed tool parameters and are not tuned on evaluation outcomes.

Table 4: **Preset ECG measurement agreement tolerances.**

Measurement type μ	Tolerance ϵ_μ
Heart rate	8 bpm
RR interval	40 ms
PR interval	30 ms
QRS duration	20 ms
QT interval	35 ms
QTc interval	40 ms
ST amplitude	0.08 mV
T-wave amplitude	0.12 mV

A.6 CLINICAL EXPERT EVALUATION PROTOCOL

Our primary evaluation relies on rubric-based LLM judges to assess diagnostic correctness and report quality. To provide a more practical clinical assessment of the actual report modifications, we additionally conduct an expert review of a sampled subset of repaired cases. The review focuses on whether the changes introduced by ECG-RePAIR are supported by the underlying ECG and improve the resulting interpretation without introducing new errors. The detailed review procedure is described below.

Case selection. We conduct a clinical expert evaluation to assess whether ECG-RePAIR's report modifications are supported by the ECG and improve the resulting interpretation. We randomly sample 50 cases without replacement from the 550 reports receiving diagnostic repair, using a fixed seed of 2027. Selection is independent of automated evaluation scores and waveform appearance. The sampled cases cover multiple repair types, including PR prolongation, axis deviation, low voltage, poor R-wave progression, and bradycardia. Each case is evaluated as a whole, including all report changes made by ECG-RePAIR.

Review materials. Each case is presented in an offline review interface containing the original ECG, the reports before and after repair, and highlighted textual changes. Experts can inspect the full ECG and individual-lead waveforms and expand both reports when needed. Model identities, reference reports, automated scores, and repair-advantage estimates are hidden during review, while the before/after order remains visible.

Assessment criteria. Two clinicians independently assess each case using the form shown in Figure 7. They judge whether the modified claims are supported by the ECG, whether the repair improves the report, and whether it introduces a new error. Reviewers also record the relevant leads, measurements or criteria, together with a brief explanation supporting their assessment. When multiple edits occur in the same case, the changes are evaluated jointly, and any conflicting judgments are documented in the explanation.

Clinician Review Form

Case ID _____ **Reviewer ID** _____

Assess the changes against the ECG. Select one response per question.

1. Claim support
Are the modified clinical assertions supported by the ECG?

- ☐ **Supported**: all modified assertions are supported.
- ☐ **Unsupported**: at least one modified assertion is unsupported.
- ☐ **Insufficient evidence**: none is clearly unsupported, but at least one cannot be assessed.

2. Clinical effect
Considering all changes, how does the repair affect the report?

- ☐ Improves ☐ No meaningful improvement
- ☐ Worsens ☐ Cannot judge

3. New error
Does the modification introduce an error?

- ☐ Yes ☐ No ☐ Cannot judge

Affected claim, if yes: _____

Supporting explanation

Relevant leads: _____

Criterion or measurement (optional): _____

Brief reason (required): _____

_____

Explain uncertain judgements and any differing assessments across edits.

Figure 7: **Assessment form used for clinical expert review.**

Independent review. Both clinicians complete their assessments independently for all 50 cases. Their individual responses are retained for analysis, allowing inter-expert agreement and disagreement to be examined without replacing them with a consensus label.

B Additional Experimental Results

This section provides further analyses of ECG-RePAIR, including robustness to the base model and evaluator, repair behavior across claim types, and runtime overhead.

B.1 Analysis of Robustness Results

Alternative base ECG model. We evaluate ECG-RePAIR with GEM as the base ECG model while keeping the same verification-and-repair framework and evaluation protocol. To compare repair effects across base models, Figure 8 plots the before–after score difference for each evaluation metric under the primary judge, separately for ECG-R1 and GEM. ECG-RePAIR improves GEM across all eight dimensions. In particular, Diagnosis Accuracy increases from 65.88 to 67.85, with additional gains in report quality, grounding, reasoning, and diagnostic fidelity. These results show that the benefit of evidence-grounded verification and repair extends beyond a single base ECG model.

The gains on GEM are smaller than those on ECG-R1, suggesting that repair effectiveness also depends on the characteristics of the initial report and on whether its remaining errors are amenable to local repair. Report-level analysis provides further context: ECG-RePAIR modifies 21.1% of GEM reports, within which Diagnosis Accuracy improves by an average of 9.34 points, yielding a 1.97-point gain over the full cohort. Thus, the smaller aggregate gain coexists with a substantially larger benefit among the reports that are actually modified, consistent with the localized design of ECG-RePAIR.

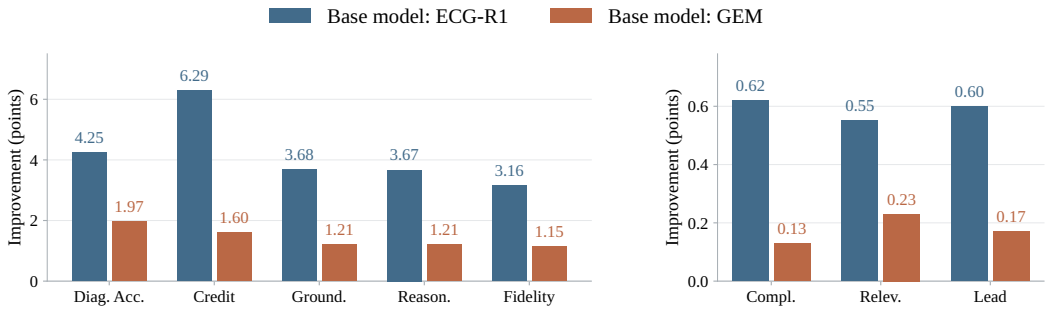


Figure 8: Improvements from ECG-RePAIR with ECG-R1 and GEM as the base model, evaluated by the primary judge. Bars show the difference between the displayed scores before and after repair. Metrics with different scales are shown separately.

Alternative evaluator. We assess evaluator sensitivity by rescoring the same stored reports with `Qwen3.8-Flash-Next-FP8`, while keeping the report outputs and evaluation rubric fixed. Table 5 reports the full results under this secondary judge. ECG-RePAIR improves over ECG-R1 across all eight dimensions, with a Diagnosis Accuracy gain of approximately 4.2 points under both Qwen and the primary DeepSeek judge.

To compare how the two evaluators respond to the same repairs, Figure 9 plots the before–after score difference for each metric under DeepSeek and Qwen, computed from the same stored base and repaired reports. As shown in the figure, Qwen assigns smaller gains in completeness, relevance, and fidelity, but a larger gain in grounding. This indicates that the two evaluators recognize the overall benefit of repair consistently, while differing in how much credit they assign to specific aspects of the repaired report.

For completeness, Qwen increases the score less often and by smaller amounts than DeepSeek. For fidelity, Qwen already assigns relatively high scores to many initial reports, leaving less room for further improvement. Grounding shows the opposite pattern, with the added waveform measurements and lead-specific evidence receiving stronger net credit under Qwen. Taken together, these results suggest that evaluator choice mainly affects the magnitude of report-quality gains rather than the overall direction of improvement.

Table 5: **Evaluation with Qwen3.8-Flash-Next-FP8 as the secondary judge.** Report outputs and the evaluation rubric are kept fixed.

Model	Diag. Acc.	Credit	Compl.	Relev.	Lead	Ground.	Reason.	Fidelity
Proprietary MLLMs								
Gemini-3.8-Flash	42.71	38.24	4.50	2.95	2.72	60.28	52.55	69.55
GPT-5.6-Sol	43.14	37.00	4.09	2.88	3.30	57.44	54.42	70.00
Open-source MLLMs								
Qwen3-VL-8B-Instruct	7.35	6.26	1.40	0.74	0.54	17.60	13.33	24.23
Gemma-4-26B-A4B-it	5.86	4.74	3.10	0.51	0.40	24.23	16.80	44.96
MiniCPM-V-4.6	14.28	12.12	1.38	1.07	0.54	22.63	19.68	29.12
InternVL3-8B-Instruct	22.08	19.37	2.16	1.46	1.88	22.52	19.49	27.43
GLM-4.1V-9B-Base	15.61	13.90	5.68	1.50	2.11	44.07	32.32	51.44
MiMo-VL-7B-SFT	21.45	18.09	4.05	2.08	0.87	36.96	30.95	50.85
Medical MLLMs								
Chiron-o1-8B	14.03	11.73	1.52	1.13	0.26	17.32	15.33	20.57
HuatuoGPT-Vision-7B	9.51	7.63	3.04	1.40	0.70	36.99	33.84	43.80
MedVLM-R1	30.26	26.46	0.91	0.63	0.70	14.30	12.50	18.45
QoQ-Med-VL-7B	30.47	27.45	3.48	2.42	1.27	35.33	32.36	48.80
ECG-specialized MLLMs								
PULSE-7B	64.51	58.34	2.10	2.97	0.37	40.00	41.30	43.85
GEM	66.47	61.12	4.28	4.79	4.92	67.59	65.71	70.25
ECG-R1-8B-RL	65.52	61.59	4.63	5.03	4.40	72.52	70.25	80.70
+ECG-RePAIR (Ours)	**69.70**	**66.07**	**4.74**	**5.25**	**4.95**	**76.57**	**73.64**	**82.02**

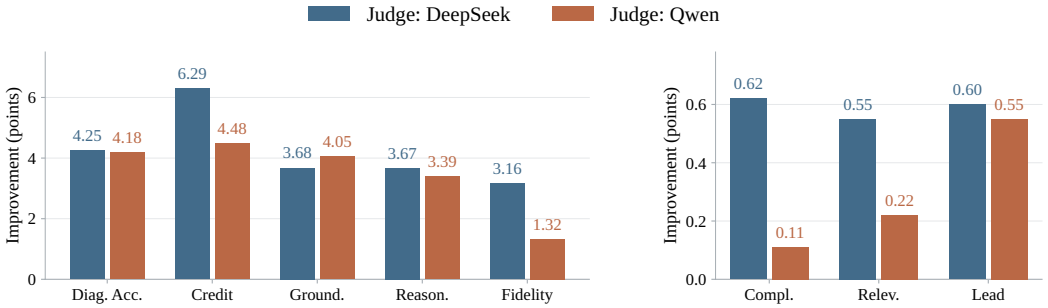


Figure 9: **Repair gains under DeepSeek and Qwen.** Bars show the before–after score differences computed from the same stored base and repaired reports. Metrics with different scales are shown separately.

B.2 Repair Patterns

To better understand how ECG-RePAIR operates in practice, we analyze how waveform verification is converted into report actions and how subsequent memory-conditioned selection further controls whether an eligible repair is actually executed. This analysis helps illustrate the role of the sequential evidence-to-action pipeline beyond aggregate performance metrics.

As shown in Figure 10, Panel (a) summarizes 35,700 registered finding instances. Interval-related findings have the highest Uncertain rate (30.2%), showing that waveform evidence is not always sufficient for a definitive judgment. This supports the need for explicit three-way verification, which allows ECG-RePAIR to preserve uncertainty rather than forcing every inspected claim into a binary decision.

Panel (b) shows that Add is the dominant repair action, particularly for voltage or morphology, axis, and interval findings. Across the 550 changed reports, ECG-RePAIR records 591 Add actions and 2 Revise actions, together with 171 explicit Keep decisions. This distribution shows that verification is translated into localized, claim-specific actions rather than unrestricted report rewriting.

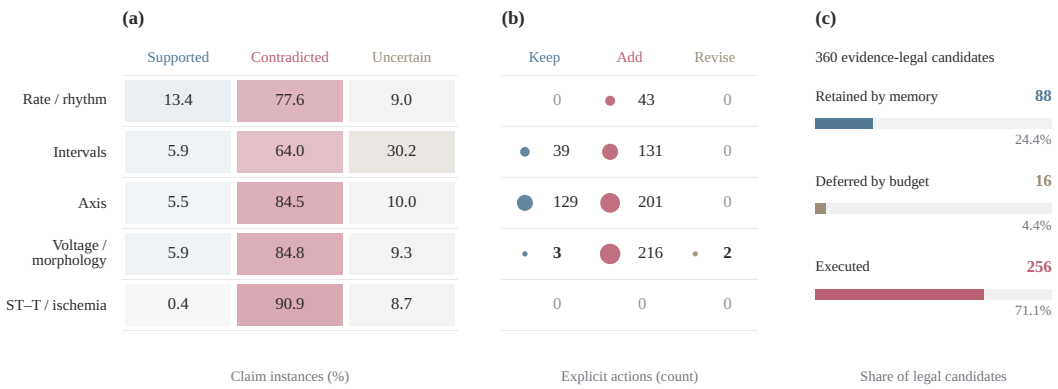


Figure 10: **Verification and repair patterns across claim types.** **(a)** Three-way verification outcomes. **(b)** Recorded repair actions by claim type. **(c)** Final decisions for candidates passing the evidence-based action filter.

Panel (c) further shows why evidence verification alone is insufficient for repair. Of the 360 candidates passing the evidence-based action filter, 88 (24.4%) are retained as KEEP after memory evaluation, 16 (4.4%) are not selected under the action budget, and 256 (71.1%) are executed. Thus, satisfying a waveform criterion only makes an action admissible; the final decision still depends on whether prior repair experience supports taking that action. Together, these patterns illustrate the complementary roles of verification, evidence-constrained action filtering, and memory-conditioned selection in preventing uncertain or weakly beneficial edits from being applied automatically.

B.3 RUNTIME OVERHEAD

Base-report generation is timed on a single NVIDIA RTX 6000 Ada Generation GPU with batch size 1. CPU repair is benchmarked on an Intel Xeon Platinum 8468 using one process and one numerical-library thread, with 50 ECGs, ten per fold, and three repetitions after warm-up. Each repair run recomputes waveform measurements and executes the repair procedure. Table 6 reports generation and repair times separately; model loading, index construction, and offline memory preparation are excluded from the timed operations.

ECG-RePAIR adds a median latency of approximately 0.162 s per ECG, with the 95th percentile below 0.179 s. Most time is spent on waveform measurement for the current ECG; verification, memory-conditioned repair, and output checks require only a few additional milliseconds. The framework therefore provides evidence-based report correction with little additional latency, without requiring another round of language-model generation or additional GPU computation during the repair process.

Table 6: Per-ECG runtime. Initial report generation uses a single NVIDIA RTX 6000 Ada Generation GPU with batch size 1. GPU results currently use one completed pass over 50 ECGs; CPU results use three repetitions per ECG. Model loading is excluded, and repair totals are timed directly.

Operation	Device	Median (ms)	95th pct. (ms)
Initial report generation (one ECG)	GPU	22,465.89	25,212.74
ECG measurement	CPU	159.35	175.57
Claim verification and report repair	CPU	1.79	5.47
Evidence graph and output checks	CPU	0.24	0.26
Total repair time	CPU	**162.03**	**178.60**
Criterion retrieval (optional)	CPU	0.33	0.64
Total with criterion retrieval	CPU	162.38	178.63

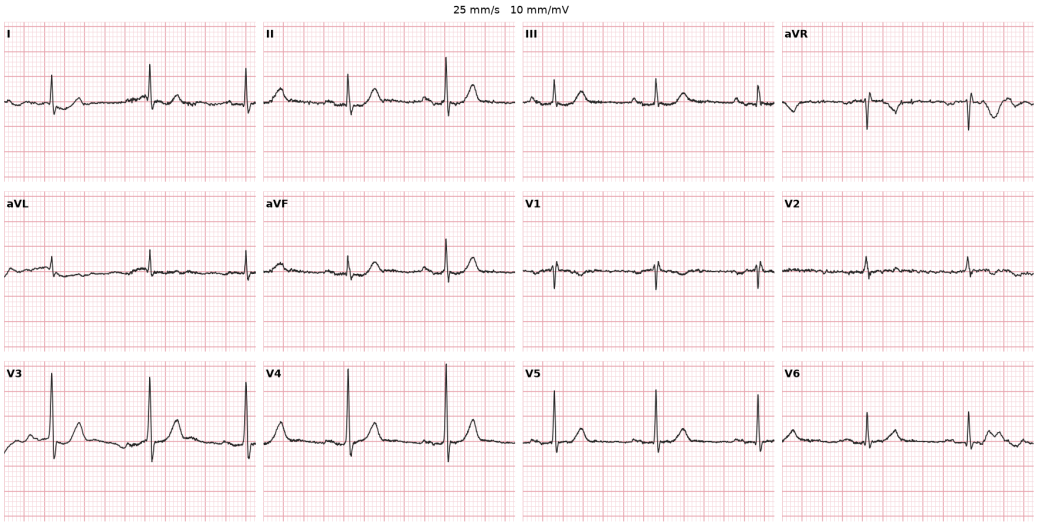
C Case Study

These two cases extend the examples in Figure 1. For each case, we first show the full 12-lead input image, followed by a lead-specific enlargement, the claim under review, the criterion and measured evidence, the recorded decision, and the complete reports before and after repair. Enlargements are drawn from the same digital recording and supplement, rather than replace, the full ECG input. Criterion cards below summarize the executable rules and are not saved retrieval transcripts. Report wording and numerical decision records come from the archived run. Reference reports and case-level judge scores are not used as inputs to the illustrated process.

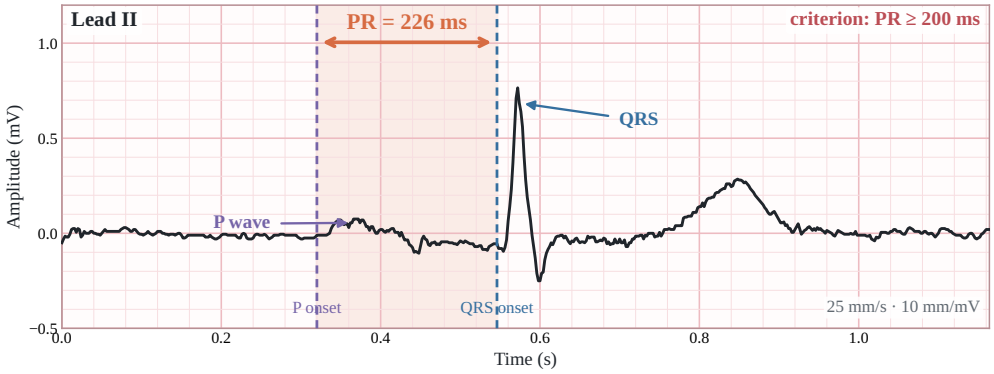
C.1 Case 1: Missed conduction abnormality

This case follows a missed conduction finding from the initial report to the final diagnostic answer. Each box records an intermediate output; the intervening text explains how that output is used in the next operation. The full ECG provides the case context, while the enlargement below illustrates the interval inspected during verification. It is not a separate input or the sole basis of the record-level measurement.

Input. Full 12-lead ECG image and corresponding digital waveform.



Lead-specific enlargement.



Lead II shows the P-wave onset and QRS onset for one representative beat. The archived record-level PR value is 225.5 ms, rendered as 226 ms.

From the base report to a claim for inspection. The base model first produces an interpretation and a diagnostic answer. The claim scanner identifies the statement about a normal PR interval as an assertion that can be checked against the current waveform. Its reported value of approximately 180 ms is retained as part of the base assertion, not accepted as a tool measurement. At this point, PR prolongation is still unverified; the agent has a claim to inspect, not yet a reason to change the report.

Claim selected from the base report

The PR interval is within normal limits at approximately 180 ms, ruling out AV block or pre-excitation syndromes.

From the selected claim to an evidence plan. The selected claim determines which criterion is needed. The criterion card specifies the measurement, lead, and conditions required to check PR prolongation. These fields become an evidence plan: locate the relevant atrial and ventricular landmarks, obtain the PR interval from the digital waveform, and check measurement agreement. The criterion provides the rule for interpreting a measurement; it does not supply the patient's measurement.

Criterion card and evidence plan

Target: PR prolongation. **Lead:** II. **Measure:** P onset to QRS onset, in milliseconds. **Rule used by the stored implementation:** PR $\geq$ 200 ms with measurable atrial activity. **Plan:** inspect the interval evidence and its cross-backend agreement before considering a diagnostic addition.

From waveform measurements to a verification verdict. The ECG tools provide the interval evidence requested by the plan. The recorded PR value is 225.5 ms, with an 11.0 ms spread between the measurement paths. This spread is within the preset PR agreement tolerance of 30 ms. The evidence card retains the value, lead, agreement result, and measurement provenance so that the verifier can apply the criterion to a traceable observation. The supported finding is PR prolongation; equivalently, the base report's opposing normal-PR assertion is contradicted.

Evidence card and verification

PR: 225.5 ms (reported as 226 ms). **Lead:** II. **Agreement:** accepted; recorded backend spread 11.0 ms. **Provenance:** interval tool, cached waveform measurement. The evidence card supports PR prolongation; the original normal-PR assertion is therefore contradicted.

Recorded decision and local execution. The verification result establishes that the proposed finding has waveform support, but does not itself execute a report edit. The updated claim state and evidence are passed to the memory-conditioned decision process to evaluate the candidate action. This separates the question of whether PR prolongation is supported from the question of whether its addition should be accepted.

Memory-conditioned decision

The logged candidate is the abnormal finding, so its state changes from UNVERIFIED to SUPPORTED. This is consistent with calling the opposing normality assertion CONTRADICTED in Figure 1. The memory-conditioned terminal decision accepts ADD, with a recorded decision value of 31.94. This is a policy value, not a measured accuracy improvement for this patient. The core trace records an ADD decision; replacing the associated report sentence is part of its execution, not a separately logged REVISE decision. No expansion action is taken.

The compact trace below places this accepted action alongside the other inspected candidates. Left-axis deviation, low voltage, and right-axis deviation are not supported in this trace and receive no diagnostic addition. Thus, inspection does not imply execution. The operation names reproduce the archived trace; its `abstain` entries indicate that no repair was committed for those candidates.

Candidate	Recorded operation	Outcome
axis left	inspect	contradicted
axis left	abstain	abstain
low voltage	inspect	contradicted
low voltage	abstain	abstain
pr prolonged	inspect	supported
pr prolonged	commit add	add
axis right	inspect	contradicted
axis right	abstain	abstain

Repeated memory lookups are omitted from this compact trace. Both cases are query-fold 0 records; the recorded memory queries use folds 1–4.

From the accepted action to the repaired report. The accepted ADD decision is then passed to the local executor. Execution has two linked effects: the targeted normal-PR sentence is replaced by the measured interval statement, and first-degree AV block is added to the diagnostic answer. The pre-existing sinus-rhythm diagnosis is retained. The subsequent evidence-rendering pass attaches supporting information without making another diagnostic decision.

Local edit and diagnostic output

The executor removes the targeted normal-PR sentence, inserts the measured PR statement, and adds first-degree AV block to the diagnostic answer. Sinus rhythm is retained. The later evidence-rendering pass does not change this diagnostic answer.
Before: Sinus rhythm
After: Sinus rhythm First degree av block (PR 226 ms)

Scope of this case

The complete archived output still contains broader statements such as 'no evidence of conduction abnormalities' in its final reasoning. This case demonstrates correction of the targeted PR statement and diagnostic answer, not complete removal of contradictory wording throughout the report.

Complete reports. The following boxes retain the complete archived text. Formatting tags are rendered as headings and emphasis; the targeted base assertion is red and the inserted measurement sentence is green.

ECG-R1: complete base report

Step 1: Technical, Rate & Rhythm The ECG demonstrates good technical quality with a stable baseline, standard 25 mm/s speed, and 10 mm/mV calibration, with no evidence of lead reversal as P waves are upright in lead I. The heart rate is 59 bpm, consistent with sinus bradycardia. The rhythm is sinus, confirmed by the presence of normal P waves that are positive in leads II, III, and aVF with a consistent PR interval; there is no irregularity to suggest atrial fibrillation or flutter, and the absence of ectopic beats or grouped beating rules out significant AV block.

Step 2: Conduction, Axis & Intervals The frontal plane axis is normal, as evidenced by positive QRS complexes in both leads I and II. The PR interval is within normal limits at approximately 180 ms, ruling out AV block or pre-excitation syndromes. The QRS duration is normal at less than 100 ms, which excludes bundle branch blocks; specifically, the absence of an rSR' pattern in V1 and a broad monophasic R wave in V6 rules out right and left bundle branch block, respectively.

Step 3: Chamber Hypertrophy & Voltage There is no evidence of left ventricular hypertrophy, as the S wave in V1 and R wave in V6 do not meet the Sokolow-Lyon voltage criteria, and the R wave in aVL is less than 11 mm. R wave progression is normal from V1 to V4, with a transition zone around V3-V4, which rules out poor R wave progression or reversed progression suggestive of right ventricular hypertrophy or posterior infarction.

Step 4: Ischemia, Infarction & Mimics A systematic analysis of all lead groups reveals no evidence of acute ischemia or infarction. There is no significant ST segment elevation or depression in the inferior leads (II, III, aVF), lateral leads (I, aVL, V5, V6), or anterior leads (V1-V4). The ST segments are predominantly isoelectric or show normal upsloping morphology. There are no pathologic Q waves in any territory, and lead aVR shows no ST elevation to suggest left main disease, effectively ruling out acute coronary syndromes.

Step 5: Electrolytes & QT The T waves have a normal morphology without tenting to suggest hyperkalemia or flattening to suggest hypokalemia. The QTc interval is within normal limits, calculated at approximately 420 ms, which is below the 440 ms threshold for prolonged QT.

Step 6: Final Medical Reasoning The ECG is interpreted as a normal tracing with sinus bradycardia. All measured intervals, axes, and waveforms are within normal limits, and there is no evidence of conduction abnormalities, chamber enlargement, acute ischemia, infarction, or electrolyte disturbances.

This ECG demonstrates sinus bradycardia at a rate of 59 bpm with normal technical quality. The rhythm is sinus, with a normal PR interval and QRS duration, confirming the absence of conduction delays or blocks. The electrical axis is normal, and there is no evidence of chamber hypertrophy based on voltage criteria. A comprehensive lead-by-lead analysis reveals no ST segment deviations, T wave abnormalities, or pathologic Q waves to suggest acute ischemia, infarction, or electrolyte imbalance. The QTc interval is normal. The overall impression is of a normal ECG variant.

Diagnostic answer Sinus rhythm

ECG-RePAIR: complete final report

Step 1: Technical, Rate & Rhythm The ECG demonstrates good technical quality with a stable baseline, standard 25 mm/s speed, and 10 mm/mV calibration, with no evidence of lead reversal as P waves are upright in lead I. The heart rate is 59 bpm, consistent with sinus bradycardia. The rhythm is sinus, confirmed by the presence of normal P waves that are positive in leads II, III, and aVF with a consistent PR interval; there is no irregularity to suggest atrial fibrillation or flutter, and the absence of ectopic beats or grouped beating rules out significant AV block.

Step 2: Conduction, Axis & Intervals The frontal plane axis is normal, as evidenced by positive QRS complexes in both leads I and II. The QRS duration is normal at less than 100 ms, which excludes bundle branch blocks; specifically, the absence of an rSR' pattern in V1 and a broad monophasic R wave in V6 rules out right and left bundle branch block, respectively. The PR interval measures 226 ms with P waves preceding each QRS complex, above the 200 ms upper limit of normal.

Step 3: Chamber Hypertrophy & Voltage There is no evidence of left ventricular hypertrophy, as the S wave in V1 and R wave in V6 do not meet the Sokolow-Lyon voltage criteria, and the R wave in aVL is less than 11 mm. R wave progression is normal from V1 to V4, with a transition zone around V3-V4, which rules out poor R wave progression or reversed progression suggestive of right ventricular hypertrophy or posterior infarction.

Step 4: Ischemia, Infarction & Mimics A systematic analysis of all lead groups reveals no evidence of acute ischemia or infarction. There is no significant ST segment elevation or depression in the inferior leads (II, III, aVF), lateral leads (I, aVL, V5, V6), or anterior leads (V1-V4). The ST segments are predominantly isoelectric or show normal upsloping morphology. There are no pathologic Q waves in any territory, and lead aVR shows no ST elevation to suggest left main disease, effectively ruling out acute coronary syndromes.

Step 5: Electrolytes & QT The T waves have a normal morphology without tenting to suggest hyperkalemia or flattening to suggest hypokalemia. The QTc interval is within normal limits, calculated at approximately 420 ms, which is below the 440 ms threshold for prolonged QT.

Step 6: Final Medical Reasoning The ECG is interpreted as a normal tracing with sinus bradycardia. All measured intervals, axes, and waveforms are within normal limits, and there is no evidence of conduction abnormalities, chamber enlargement, acute ischemia, infarction, or electrolyte disturbances.

This ECG demonstrates sinus bradycardia at a rate of 59 bpm with normal technical quality. The rhythm is sinus, with QRS duration, confirming the absence of conduction delays or blocks. The electrical axis is normal, and there is no evidence of chamber hypertrophy based on

voltage criteria. A comprehensive lead-by-lead analysis reveals no ST segment deviations, T wave abnormalities, or pathologic Q waves to suggest acute ischemia, infarction, or electrolyte imbalance. The QTc interval is normal. The overall impression is of a normal ECG variant. The PR interval measures 226 ms with P waves preceding each QRS complex, above the 200 ms upper limit of normal. There is first degree AV block.

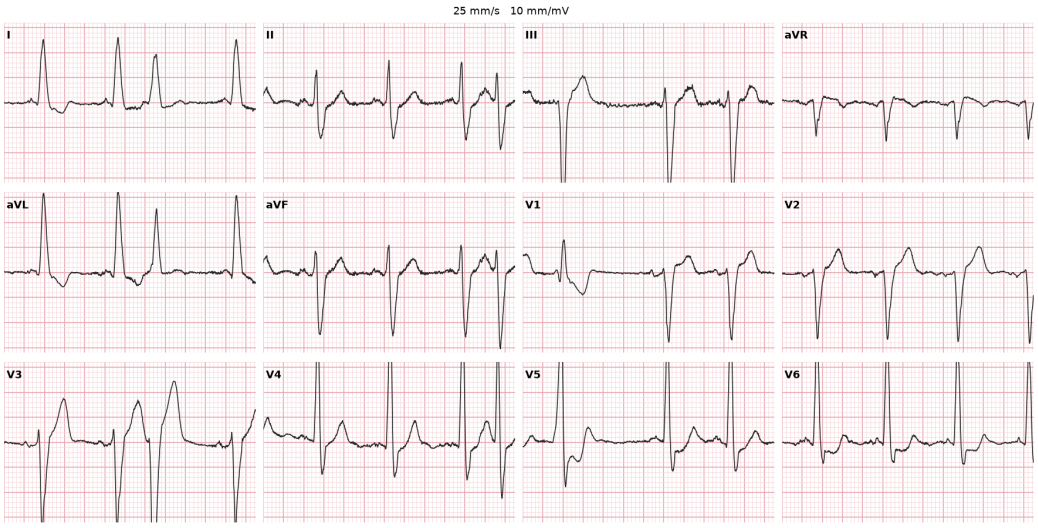
Diagnosis-linked ECG evidence verified by the ECG Agent: 1. **sinus rhythm.** ventricular rate 60 bpm; lead-II RR coefficient of variation 0.026 (below the 0.12 marked-irregularity gate); lead-II pre-QRS P-wave presence fraction 1.00 with measurable atrial morphology. Leads assessed: II. 2. **first degree av block.** PR 226 ms; meeting the >=200 ms first-degree AV-block criterion. Leads assessed: II.

Diagnostic answer Sinus rhythm First degree av block (PR 226 ms)

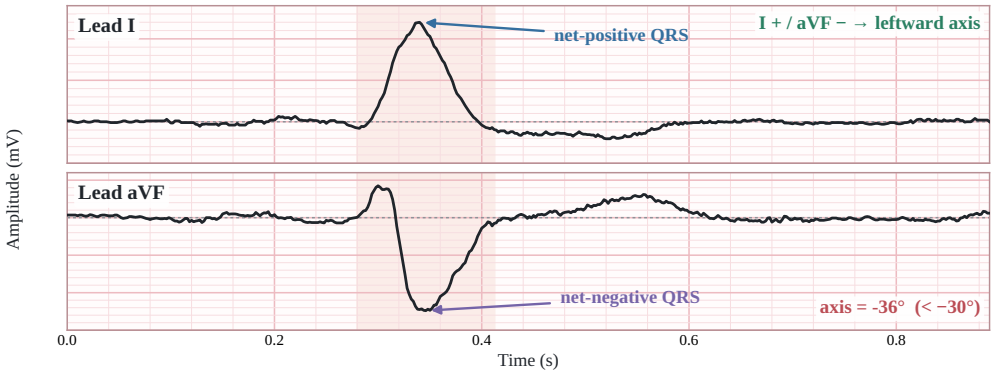
C.2 Case 2: Contradicted axis claim

This case follows the same procedure for an axis claim. The base report contains several diagnoses, but the repair considered here targets its normal-axis assertion. The full ECG is shown first; the enlarged leads then identify the waveform evidence used for this specific verification task.

Input. Full 12-lead ECG image and corresponding digital waveform.



Lead-specific enlargement.



Leads I and aVF show the net QRS deflections used for the frontal-axis estimate. The archived axis is -35.8 degrees, rendered as -36 degrees.

From the base report to a claim for inspection. The claim scanner identifies the normal-axis statement and its lead-polarity justification in the initial report. This provides both the assertion to check and the report span that may require modification. The polarity statement is still a model assertion, not a verified observation. Left-axis deviation remains unverified until the corresponding waveform evidence is obtained.

Claim selected from the base report

The frontal plane axis is normal, as leads I and II are both positive.

From the selected claim to an evidence plan. The axis criterion specifies the angular range associated with left-axis deviation and the leads used by the measurement tool. The resulting plan requests a frontal-axis estimate from the net QRS deflections in leads I and aVF, together with an agreement check. This converts a qualitative normal-axis claim into a measurement-based verification task.

Criterion card and evidence plan

Target: left-axis deviation. **Leads:** I and aVF. **Measure:** frontal QRS axis from net QRS deflections. **Rule:** $-90^\circ \leq \theta < -30^\circ$. **Plan:** inspect the lead-specific axis estimate and detector agreement; a verbal lead-polarity assertion alone is not sufficient.

From waveform measurements to a verification verdict. The tool returns an axis estimate of -35.8° and records agreement between the two detector paths. The evidence card associates this result with leads I and aVF and the net-QRS measurement procedure. Applying the retrieved criterion supports left-axis deviation and contradicts the normal-axis conclusion. This verdict concerns the axis claim only; it does not establish the validity of every diagnosis in the initial report.

Evidence card and verification

Axis: -35.8° (reported as -36°). **Leads:** I and aVF. **Agreement:** two detector paths, 0.0° spread. **Provenance:** median net-QRS deflection tool. The evidence card supports left-axis deviation; the original normal-axis assertion is contradicted.

Recorded decision and local execution.

The supported axis finding and its evidence next become inputs to memory-conditioned action selection. The recorded decision accepts the diagnostic addition. As in the first case, this is a separate step from verification: waveform evidence establishes support for the finding, while the decision process determines whether the candidate repair is executed.

Memory-conditioned decision

The logged candidate is the abnormal finding, so its state changes from Unverified to Supported. This is consistent with calling the opposing normality assertion Contradicted in Figure 1. The memory-conditioned terminal decision accepts Add, with a recorded decision value of 89.91. This is a policy value, not a measured accuracy improvement for this patient. The core trace records an Add decision; replacing the associated report sentence is part of its execution, not a separately logged Revise decision. No expansion action is taken.

The trace below shows the distinction between an accepted candidate and an inspected alternative. Left-axis deviation is supported and committed as an addition; right-axis deviation is contradicted and receives no edit. The `abstain` entry is the archived operation name for this non-executed candidate.

Candidate	Recorded operation	Outcome
axis left	inspect	supported
axis left	commit add	add
axis right	inspect	contradicted
axis right	abstain	abstain

Repeated memory lookups are omitted from this compact trace. Both cases are query-fold 0 records; the recorded memory queries use folds 1–4.

From the accepted action to the repaired report. The executor uses the accepted action to update the targeted axis statement and append left-axis deviation to the diagnostic answer. The final evidence-rendering pass presents the measured angle and its lead-specific support. Other diagnoses are carried forward rather than re-derived by this axis repair. The before-and-after output below therefore separates the newly accepted finding from the pre-existing diagnostic content.

Local edit and diagnostic output

The executor removes the normal-axis conclusion, inserts the measured axis statement, and adds left-axis deviation to the diagnostic answer. Other diagnoses remain unchanged. The final rendering attaches axis evidence without adding further diagnoses.
Before: Sinus rhythm; Premature ventricular complexes; Premature atrial complexes; Left bundle branch block; Left ventricular hypertrophy
After: Sinus rhythm; Premature ventricular complexes; Premature atrial complexes; Left bundle branch block; Left ventricular hypertrophy Left axis deviation (frontal axis -36 degrees)

Scope of this case

This is an axis-specific repair, not independent confirmation of the retained PVC, PAC, LBBB, or LVH diagnoses. The archived output also retains the original statement that leads I and II are positive. We reproduce it rather than silently correcting the model's remaining text. The displayed lead crop should not be used to validate the entire diagnostic list.

Complete reports. The following boxes retain the complete archived text. Formatting tags are rendered as headings and emphasis; the targeted base assertion is red and the inserted measurement sentence is green.

ECG-R1: complete base report

Step 1: Technical, Rate & Rhythm The ECG demonstrates good technical quality with a stable baseline, standard 25 mm/s speed, and 10 mm/mV calibration, with no evidence of lead reversal as P waves are upright in lead I. The rhythm is sinus with a rate of 84 bpm, confirmed by the presence of positive P waves in leads II, III, and aVF. The rhythm is irregular due to the presence of premature ventricular complexes (PVCs) and premature atrial complexes (PACs), as evidenced by the variable RR intervals and the presence of both normal and aberrant QRS complexes, ruling out atrial fibrillation or flutter which lack organized P waves.

Step 2: Conduction, Axis & Intervals The frontal plane axis is normal, as leads I and II are both positive. The PR interval is within normal limits at approximately 160 ms, ruling out AV block. The QRS duration is prolonged at 140 ms, indicating a complete bundle branch block. The morphology in V1 shows a broad, monophasic R wave with a slurred upstroke, and V6 demonstrates a deep, wide S wave, which is diagnostic of a left bundle branch block (LBBB), thereby ruling out a right bundle branch block pattern.

Step 3: Chamber Hypertrophy & Voltage There is evidence of left ventricular hypertrophy (LVH) based on voltage criteria, with an R wave amplitude in lead aVL measuring 11 mm and the S wave in V1 plus the R wave in V6 exceeding 35 mm. R wave progression is abnormal across the precordial leads, with poor R wave growth from V1 to V4, which is consistent with the underlying LBBB and LVH, ruling out right ventricular hypertrophy which would show a dominant R wave in V1.

Step 4: Ischemia, Infarction & Mimics A systematic analysis of all lead groups reveals no evidence of acute ischemia or infarction. There are no pathologic Q waves in the inferior (II, III, aVF), lateral (I, aVL, V5, V6), or anterior (V1-V4) leads. The ST segments and T waves show discordant changes secondary to the LBBB, with ST depression and T wave inversion in leads with a predominantly positive QRS complex (e.g., V5, V6), and ST elevation in leads with a predominantly negative QRS complex (e.g., V1), which is a normal finding in LBBB and rules out STEMI. There is no concordant ST elevation to suggest acute coronary syndrome.

Step 5: Electrolytes & QT The QT interval is prolonged with a QTc calculated at approximately 500 ms, which is abnormal. The T waves do not show the characteristic tented morphology of hyperkalemia or the flat T waves with prominent U waves of hypokalemia; the T wave morphology is consistent with the secondary repolarization abnormalities of LBBB.

Step 6: Final Medical Reasoning The ECG findings are consistent with sinus rhythm with ectopy, a complete left bundle branch block, and left ventricular hypertrophy. The prolonged QTc interval is a notable finding requiring clinical correlation. The pattern rules out acute coronary syndromes, pericarditis, and significant electrolyte disturbances based on the characteristic secondary repolarization changes of LBBB.

This ECG demonstrates sinus rhythm at 84 bpm with occasional premature ventricular and atrial complexes. The conduction analysis reveals a normal axis and PR interval but a wide QRS complex diagnostic of a complete left bundle branch block. Voltage criteria in the limb and precordial leads are met for left ventricular hypertrophy. There is no electrocardiographic evidence of acute ischemia or infarction, as the ST-T wave changes are secondary to the conduction abnormality. A prolonged QTc interval is present. The overall impression is consistent with the provided diagnoses.

Diagnostic answer Sinus rhythm; Premature ventricular complexes; Premature atrial complexes; Left bundle branch block; Left ventricular hypertrophy

ECG-RePAIR: complete final report

Step 1: Technical, Rate & Rhythm The ECG demonstrates good technical quality with a stable baseline, standard 25 mm/s speed, and 10 mm/mV calibration, with no evidence of lead reversal as P waves are upright in lead I. The rhythm is sinus with a rate of 84 bpm, confirmed by the presence of positive P waves in leads II, III, and aVF. The rhythm is irregular due to the presence of premature ventricular complexes (PVCs) and premature atrial complexes (PACs), as evidenced by the variable RR intervals and the presence of both normal and aberrant QRS complexes, ruling out atrial fibrillation or flutter which lack organized P waves.

Step 2: Conduction, Axis & Intervals Leads I and II are both positive. The PR interval is within normal limits at approximately 160 ms, ruling out AV block. The QRS duration is prolonged at 140 ms, indicating a complete bundle branch block. The morphology in V1 shows a broad, monophasic R wave with a slurred upstroke, and V6 demonstrates a deep, wide S wave, which is diagnostic of a left bundle branch block (LBBB), thereby ruling out a right bundle branch block pattern. The frontal QRS axis measures -36 degrees from the net QRS deflections in leads I and aVF, leftward of the -30 degree limit.

Step 3: Chamber Hypertrophy & Voltage There is evidence of left ventricular hypertrophy (LVH) based on voltage criteria, with an R wave amplitude in lead aVL measuring 11 mm and the S wave in V1 plus the R wave in V6 exceeding 35 mm. R wave progression is abnormal across the precordial leads, with poor R wave growth from V1 to V4, which is consistent with the underlying LBBB and LVH, ruling out right ventricular hypertrophy which would show a dominant R wave in V1.

Step 4: Ischemia, Infarction & Mimics A systematic analysis of all lead groups reveals no evidence of acute ischemia or infarction. There are no pathologic Q waves in the inferior (II, III, aVF), lateral (I, aVL, V5, V6), or anterior (V1-V4) leads. The ST segments and T waves show discordant changes secondary to the LBBB, with ST depression and T wave inversion in leads with a predominantly positive QRS complex (e.g., V5, V6), and ST elevation in leads with a predominantly negative QRS complex (e.g., V1), which is a normal finding in LBBB and rules out STEMI. There is no concordant ST elevation to suggest acute coronary syndrome.

Step 5: Electrolytes & QT The QT interval is prolonged with a QTc calculated at approximately 500 ms, which is abnormal. The T waves do not show the characteristic tented morphology of hyperkalemia or the flat T waves with prominent U waves of hypokalemia; the T wave morphology is consistent with the secondary repolarization abnormalities of LBBB.

Step 6: Final Medical Reasoning The ECG findings are consistent with sinus rhythm with ectopy, a complete left bundle branch block, and left ventricular hypertrophy. The prolonged QTc interval is a notable finding requiring clinical correlation. The pattern rules out acute coronary syndromes, pericarditis, and significant electrolyte disturbances based on the characteristic secondary repolarization changes of LBBB.

This ECG demonstrates sinus rhythm at 84 bpm with occasional premature ventricular and atrial complexes. The conduction analysis reveals PR interval but a wide QRS complex diagnostic of a complete left bundle branch block. Voltage criteria in the limb and precordial leads are met for left ventricular hypertrophy. There is no electrocardiographic evidence of acute ischemia or infarction, as the ST-T wave changes are secondary to the conduction abnormality. A prolonged QTc interval is present. The overall impression is consistent with the provided diagnoses. The frontal QRS axis measures -36 degrees from the net QRS deflections in leads I and aVF, leftward of the -30 degree limit. There is left axis deviation.

Diagnosis-linked ECG evidence verified by the ECG Agent: 1. **left axis deviation.** frontal QRS axis -36 degrees; left axis deviation by the median net-QRS vector in leads I and aVF; 2 detector paths, 0.0-degree spread. Leads assessed: I, aVF. 2. **left ventricular hypertrophy.** Sokolow-Lyon 3.86 mV; R(aVL) 1.48 mV; R(I) 1.14 mV; Sokolow-Lyon 3.86 mV >3.50 mV and R(aVL) 1.48 mV >1.10 mV support LVH by voltage. Leads assessed: I, aVL, V1, V5, V6.

Diagnostic answer Sinus rhythm; Premature ventricular complexes; Premature atrial complexes; Left bundle branch block; Left ventricular hypertrophy Left axis deviation (frontal axis -36 degrees)

C.3 Case Study Summary

The two cases illustrate different sources of error in the initial interpretation. In the first case, the base report estimates the PR interval at approximately 180 ms and concludes that conduction is normal. The measured value of 225.5 ms contradicts this assertion and supports adding first-degree AV block. In the second case, the report asserts a normal frontal axis from a verbal description of lead polarity. The measured axis of -35.8° instead supports left-axis deviation. The first correction therefore addresses an inaccurate interval estimate, whereas the second replaces an unsupported axis conclusion with an explicit measurement.

In both cases, the criterion-guided plan identifies the relevant leads and measurement, allowing verification to revisit a specific assertion rather than repeat the entire interpretation. The resulting Evidence Card makes the reason for correction inspectable: the measured value, lead set, and clinical boundary are recorded together. The memory-conditioned decision then accepts an Add action, and the executor updates the associated report sentence while including the missing finding in the diagnostic answer.

These examples show how ECG-RePAIR connects a report-level error to waveform evidence and a targeted diagnostic update. Other inspected candidates, such as right-axis deviation, are rejected when their measurements contradict the proposed finding. This makes both the accepted correction and the decision to leave other claims unchanged traceable. The complete outputs also show the limits of local repair: some broader contradictory wording remains, and retaining an unrelated diagnosis does not independently verify it. The demonstrated benefit is therefore a measurable, localized correction with an explicit evidence trail, while consistency across the full report remains an area for improvement.

D Further Discussion

Limitations. ECG-RePAIR operates after initial report generation and remains constrained by the quality of the initial interpretation and the coverage of the available verification tools. Future work could integrate evidence-to-action reasoning directly into model training.

Broader impact. ECG-RePAIR improves automated ECG interpretation by verifying uncertain findings against the waveform and making repair decisions more traceable. It is intended to support clinical review rather than replace expert interpretation, and further validation is required before clinical deployment.