ECG-EGR: EVENT-GROUNDED RELATIONAL REASONING FOR CLINICAL ECG INTERPRETATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Electrocardiography (ECG) is fundamental to cardiovascular diagnosis, and deep learning has substantially advanced automated ECG classification and report generation. However, most existing approaches encode ECGs into latent representations optimized for downstream prediction, leaving the correspondence between model representations, localized waveform evidence, and clinical findings largely implicit. We introduce ECG-EGR, an Event-Grounded Relational framework that factorizes ECG interpretation through a sparse, variable-cardinality set of waveform-supported Events. Each Event carries explicit lead-time support, an immutable semantic address, and a mutable reasoning state, separating evidence localization from contextual interpretation. ECG-EGR first discovers grounded Events and aligns their semantics with weak clinical supervision through unbalanced optimal transport, then relationally composes Event states to capture temporal and cross-lead dependencies while preserving their grounded identity. A language model accesses ECG information only through the composed Event representations, while a finding-conditioned retriever links generated findings to immutable Event semantics, enabling structured provenance from clinical findings to source Events and waveform regions. Extensive experiments demonstrate strong report generation performance, including a BLEU-1 score of 0.876, while achieving the strongest cross-task representation transfer among the compared methods.

1 INTRODUCTION

Electrocardiography (ECG) is a ubiquitous, low-cost, and non-invasive test that provides rich information for diagnosing cardiovascular abnormalities Kalmady et al. (2024). In clinical practice, ECG interpretation relies on identifying waveform abnormalities in rhythm, intervals, and morphology, and integrating lead-specific patterns and inter-lead relationships into diagnostic judgments Al-Zaiti et al. (2023). Recent ECG representation learning has substantially improved classification, transfer, and zero-shot diagnosis by learning increasingly generalizable signal representations Liu et al. (2024a). In parallel, large language and multimodal models have extended ECG analysis beyond predefined labels to report generation, question answering, and open-ended interpretation, shifting the field toward language-based clinical reasoning Liu et al. (2026).

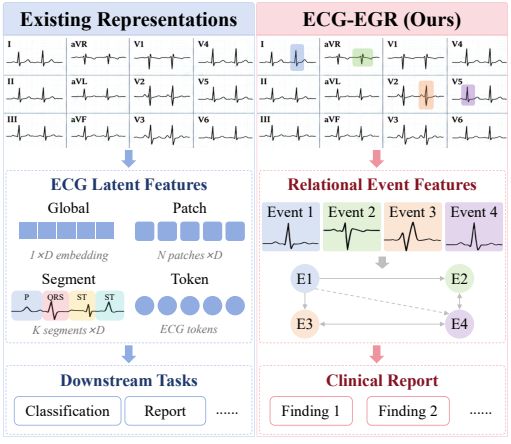


Figure 1: Comparison of existing ECG representation paradigms and ECG-EGR.

Existing ECG representations can be broadly categorized into two families. Encoder-centric approaches map recordings into global embeddings or dense temporal and lead-time features for signal modeling and downstream prediction Lai et al. (2023); Vaid et al. (2023); Na et al. (2024). Structure- and language-aware approaches instead represent ECGs using physiological segments, discrete waveform tokens, or multimodal features aligned with clinical text and LLMs Liu et al.

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.

(2024a); Jin et al. (2025); Han et al. (2025); Lan et al. (2025). Despite their success, these representations still exhibit two fundamental limitations for clinical ECG interpretation:

Predefined Granularity of ECG Representations. Many ECG encoders construct representations at a granularity determined by architecture or preprocessing, such as whole-record embeddings, uniform temporal patches, or predefined physiological segments. Yet diagnostically relevant ECG patterns are heterogeneous: morphology may be localized within a beat, rhythm abnormalities emerge over longer intervals, and informative patterns may span multiple leads and temporal scales Wang et al. (2025); McKeen et al. (2025). Although recent multi-scale approaches better capture this hierarchy, the representation structure is typically specified a priori rather than adapted to each recording.

Implicit ECG-to-Finding Reasoning. Modern ECG-language models generate reports and answer open-ended questions by projecting signal or image representations into language models Yang et al. (2025b); Wan et al. (2025); Khunte et al. (2026). These systems improve linguistic flexibility, while recent ECG-MLLMs further incorporate waveform localization and related interpretation tasks Li et al. (2026). However, the transformation from distributed ECG features to an individual clinical finding is typically carried out through latent feature interactions. Consequently, how temporal, morphological, and cross-lead information supports each finding remains difficult to inspect or trace.

To overcome these limitations, we propose **ECG-EGR**, an Event-Grounded Relational framework that reformulates ECG interpretation as reasoning over a sparse, variable-cardinality set of waveform-supported Events (see Fig 1). Each Event preserves explicit lead-time support, a stable semantic address, and a mutable reasoning state, separating evidence localization from contextual interpretation. **Competitive Event Discovery (CED)** identifies adaptive Events from the ECG while restricting their content to selected waveform support, establishing an explicit correspondence between representation and signal origin. **Compositional Evidence Transport (CET)** aligns these Events with report findings through unbalanced optimal transport, enabling clinically meaningful Event semantics under weak supervision and allowing multiple Events to jointly support a finding. Building on this representation, the **Relational Event Composer (REC)** models temporal, cross-lead, synchrony, and support relationships between Events while preserving their grounding and semantic identity. The composed Event representations are then used for report generation, and a finding-conditioned retriever links generated findings to supporting Events and waveform regions. Together, ECG-EGR provides a unified path from waveform-grounded representation to relational reasoning, report generation, and structured provenance. Our contributions are as follows:

- We introduce ECG-EGR, a new representation paradigm that organizes ECGs into sparse, variable-cardinality, waveform-grounded Events, explicitly connecting localized signal evidence with clinical reasoning and language generation.
- We develop Competitive Event Discovery and Compositional Evidence Transport to discover waveform-supported Events and ground their semantics using weak clinical supervision via unbalanced optimal transport, without Event-level annotations.
- We introduce the Relational Event Composer to model temporal, cross-lead, synchrony, and support relationships while preserving Event identity, and a finding-conditioned retriever that traces generated findings to source Events and waveform regions.
- We conduct extensive evaluations spanning report generation, cross-task transfer, zero-shot recognition, and representation stability, showing that ECG-EGR learns effective, transferable, robust, and prediction-relevant Event representations across diverse clinical settings.

2 Related Work

2.1 ECG Representation Learning

ECG representation learning has evolved from continuous global and patch-based features toward physiologically structured and language-compatible forms. Early contrastive methods learned transferable global or local embeddings, while masked and foundation models encoded recordings as dense temporal latents, image patches, or lead–time patches Kiyasseh et al. (2021); Oh et al. (2022); Lai et al. (2023); Vaid et al. (2023); Na et al. (2024); McKeen et al. (2025). More recent work incorporates cardiac structure into representation design. HeartLang models QRS-centered heartbeats

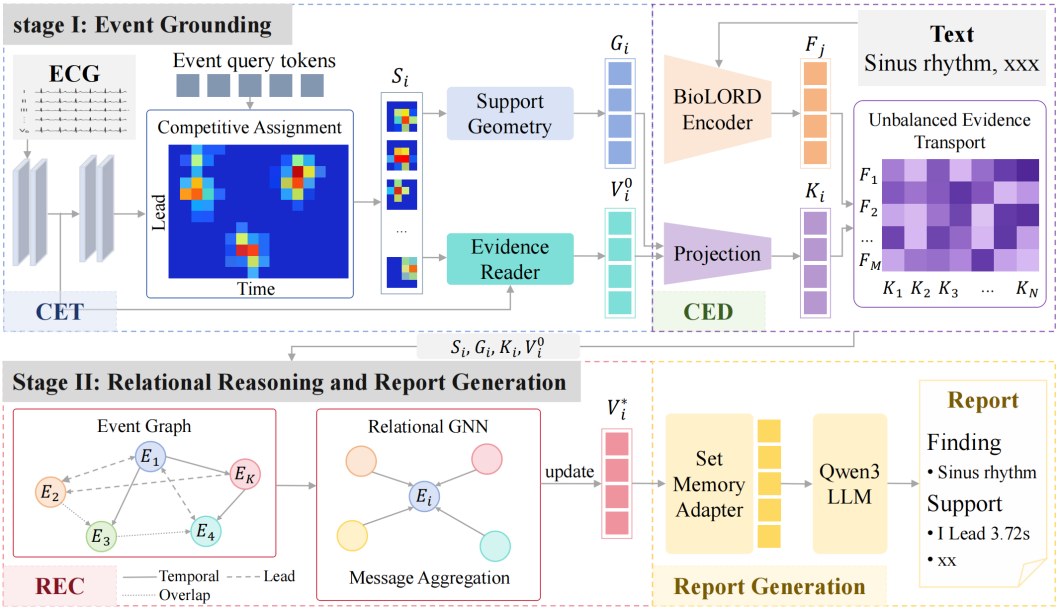


Figure 2: Overview of ECG-EGR, comprising waveform-supported Event grounding in Stage I and relational Event reasoning with structured clinical report generation in Stage II.

as words and rhythms as sentences Jin et al. (2025), whereas ECG-Byte converts quantized waveforms into variable-length discrete tokens Han et al. (2025). In parallel, multimodal approaches increasingly align ECG features with clinical language. MERL learns a shared ECG–report embedding space for zero-shot diagnosis, while D-BETA, ECG-CLIP, and ECG-LM extend this direction through masked ECG–text learning, natural-language supervision, and LLM-compatible latent representations Liu et al. (2024a); Pham Hung et al. (2025); Zhou et al. (2025); Yang et al. (2025b). Despite these advances, representation structure remains largely prescribed by encoder resolution, segmentation, or tokenization strategy rather than adapted to recording-specific diagnostic patterns.

2.2 ECG Report Generation

ECG report generation has progressed from encoder–decoder architectures to multimodal LLM-based interpretation Bartels et al. (2022); Qiu et al. (2023); Bleich et al. (2024). ECG-ReGen augments generation with retrieval from similar ECG reports Tang et al. (2025), while MEIT introduces ECG–LLM alignment and instruction tuning Wan et al. (2025). ECG-Chat and BiECG-LLM extend this direction toward conversational and bimodal ECG understanding Zhao et al. (2025); Qiang et al. (2025). Recent systems emphasize multimodal and grounded reasoning. GEM integrates ECG signals, images, and text for grounded interpretation Lan et al. (2025), whereas ECG-GPT and PULSE scale image-based report generation and instruction tuning Khunte et al. (2026); Liu et al. (2026). anyECG-chat supports waveform localization and flexible ECG inputs Li et al. (2026), while ECG-R1 introduces protocol-guided, evidence-aware reasoning Jin et al. (2026). Despite this progress, ECG report generation systems still largely rely on increasingly complex latent representations, leaving the structure of ECG interpretation itself only weakly characterized.

3 Method

In this section, we provide a detailed explanation of the overall ECG-EGR framework. Given a multi-lead ECG recording $X \in \mathbb{R}^{C\times T}$ and its paired clinical report R, ECG-EGR represents the recording as a variable-cardinality Event set $\mathcal{E}(X) = \{E_i\}_{i=1}^{N(X)}$. Each Event is defined as $E_i = (S_i, K_i, V_i)$, where S_i denotes explicit waveform support, K_i a stable semantic address, and V_i a mutable reasoning state. As illustrated in Figure 2, ECG-EGR progressively discovers, grounds, composes, and decodes Events for report generation and waveform-level provenance.

3.1 COMPETITIVE EVENT DISCOVERY

A key idea of ECG-EGR is to construct Events according to each recording's waveform structure rather than imposing a fixed waveform partition. Since clinically relevant patterns vary in temporal extent and lead distribution, we encode the ECG into two aligned feature maps. A shared lead-wise convolutional encoder extracts local morphology independently for each lead, while an axial context encoder models within-lead temporal dependencies and synchronous cross-lead interactions.

$$H_{\mathrm{local}} = f_{\mathrm{local}}(X), \qquad H_{\mathrm{ctx}} = f_{\mathrm{ctx}}(H_{\mathrm{local}}).$$

Here, H_{ctx} provides contextual information for locating relevant evidence, whereas H_{local} preserves localized waveform content from which grounded Event representations are constructed.

Competitive Support Assignment. To adapt both Event support and cardinality to each recording, we introduce $K_{\max}$ learnable Event queries. Each query interacts with H_{ctx} to produce compatibility scores z_{ilt} over lead–time locations, together with an explicit background score $z_{\mathrm{bg},lt}$ for evidence not assigned to any Event. The candidates compete locally through sparse Entmax

$$[S_{1lt}, \ldots, S_{K_{\max}lt}, S_{\mathrm{bg},lt}] = \mathrm{Entmax}_{1.5}([z_{1lt}, \ldots, z_{K_{\max}lt}, z_{\mathrm{bg},lt}]).$$

The supports satisfy $\sum_i S_{ilt} + S_{\mathrm{bg},lt} = 1$, encouraging different Events to specialize in distinct waveform regions. Queries are refined using support-weighted contextual evidence and rescored. Candidates with insufficient support or negligible relative mass are pruned, with their support returned to the background, yielding a recording-dependent Event count $N(X) \leq K_{\max}$.

Support-Constrained Evidence Reading. Localization alone does not guarantee that Event representations are grounded in the selected evidence. We therefore follow a *select with context, read from evidence* principle. Although H_{ctx} determines S_i, the initial Event state is extracted exclusively from H_{local} under this support, giving $V_i^0 = \mathcal{R}(H_{\mathrm{local}}, S_i)$. The reader aggregates support-weighted mean responses, salient activations, and local feature variation, without directly accessing H_{ctx} or the Event query. This prevents direct contextual feature access from bypassing support-constrained evidence reading. Finally, we derive $G_i = \phi(S_i)$ deterministically to summarize temporal and lead-wise support geometry for subsequent semantic alignment and relational composition.

3.2 COMPOSITIONAL EVIDENCE TRANSPORT

Semantic Event Keys. Waveform grounding identifies where an Event is supported, but its support alone does not determine what the Event means clinically. We therefore associate each Event with a semantic key that remains stable while its reasoning state is updated. The key is constructed exclusively from ECG-derived information by combining the initial Event state with support geometry, $K_i = \mathrm{Norm}\left(g_V(V_i^0) + \alpha_G g_G(G_i^{\mathrm{sem}})\right)$. Here, g_V and g_G are learned projections, and G_i^{sem} contains geometry descriptors for semantic alignment. Each report finding f is encoded by a frozen BioLORD encoder as Z_f. These report embeddings supervise the Event semantic space without participating in the construction of K_i. The resulting key therefore serves as a stable ECG-derived semantic address that remains unchanged during relational composition.

Unbalanced Evidence Transport. A one-to-one correspondence is overly restrictive for ECG interpretation because a clinical finding may depend on several Events, while some waveform evidence may remain unmentioned in the report. We therefore model finding–Event alignment as compositional transport. Given the semantic cost $C_{fi} = 1 - \cos(Z_f, K_i)$, finding mass a, and support-derived Event capacity b, we define the unbalanced optimal transport objective as

$$\mathcal{L}_{\mathrm{UOT}} = \min_{\Gamma \geq 0} \Big\{ \langle \Gamma, C \rangle + \epsilon\, \mathcal{H}(\Gamma) + \rho_f\, \mathrm{KL}(\Gamma \mathbf{1} \,\|\, a) + \rho_e\, \mathrm{KL}(\Gamma^\top \mathbf{1} \,\|\, b) \Big\}.$$

Here, Γ distributes finding mass across grounded Events. We use $\rho_f > \rho_e$ to encourage finding coverage while allowing unreported Events to receive little transport mass. The optimal plan Γ^* captures a soft many-to-many correspondence rather than forcing discrete assignments. For each finding with positive transported mass, we define $Q_{fi} = \Gamma^*_{fi} / \sum_j \Gamma^*_{fj}$, yielding a distribution over supporting Events retained as weak supervision for finding-conditioned Event retrieval.

3.3 Relational Event Composer

Relation Graph. Individual Events capture waveform-grounded evidence, yet many ECG findings depend on how patterns interact across time and leads. We therefore construct an Event graph whose topology is determined solely by waveform-derived support and geometry, avoiding semantic labels or report information when defining inter-Event structure. For each pair of Events, we compute a relation descriptor $R_{ij} = \rho(S_i, G_i, S_j, G_j)$, where $\rho(\cdot)$ summarizes temporal proximity, cross-lead correspondence, synchrony, and support overlap. These relations characterize how two Events are positioned with respect to one another rather than what diagnosis they may represent. Geometry-based affinities are then used to retain a sparse neighborhood for each Event, yielding a compact graph that captures the structural organization of evidence within the recording.

Event Composition. Once the relation graph is established, REC composes Event information while preserving the grounding and semantic identity learned earlier. Support geometry is incorporated into each mutable state, after which relation-aware attention aggregates neighboring Event states using content and pairwise relation descriptors. A gated residual update produces the composed state V_i^*, yielding $V_i^0 \to V_i^*$, while S_i, G_i, and K_i remain unchanged during composition. This separation allows contextual reasoning to modify how an Event is interpreted without altering its original support or semantic address. REC additionally maintains a detached source-Event lineage matrix A_{REC}, which summarizes information routing from source Events to composed Events and is later used to construct finding-level waveform provenance.

3.4 Event-Grounded Report Generation

Event Set Memory and Finding-First Generation. After relational composition, the variable-cardinality Event set provides the sole ECG representation to the Qwen3 Yang et al. (2025a) model. To obtain a fixed-length interface without imposing arbitrary Event ordering, we use a permutation-invariant set adapter to transform the composed states $\{V_i^*\}_{i=1}^{N(X)}$ into continuous memory M. No positional embedding is introduced along the Event dimension, preserving the set structure learned earlier. Conditioned on this memory, the language model generates structured clinical findings autoregressively in a clinically coherent manner. Each finding is completed before its supporting Events are retrieved, allowing evidence retrieval to condition on the generated clinical content. This finding-first formulation separates linguistic realization from subsequent evidence citation.

Dynamic Event Retrieval and Waveform Provenance. Once a finding f is generated, the hidden state at its citation position is projected into the Event semantic space to obtain a query q_f. Supporting Events are then dynamically retrieved through a content-aware matching mechanism:

$$P_{\mathrm{comp}}(i \mid f) = \frac{\exp\left(\cos(q_f, K_i)/\tau\right)}{\sum_{j=1}^{N(X)} \exp\left(\cos(q_f, K_j)/\tau\right)},$$

where the immutable keys K_i provide sample-specific semantic addresses rather than fixed Event identifiers, and normalization is restricted to the active Events of the current recording. Because each composed Event may inherit information from multiple source Events, the retrieval distribution is propagated through the REC lineage as $P_{\mathrm{src}}(\cdot \mid f) = P_{\mathrm{comp}}(\cdot \mid f) A_{\mathrm{REC}}$. The resulting source weights are finally projected onto waveform support through $S_f(l,t) = \sum_i P_{\mathrm{src}}(i \mid f) S_i(l,t)$, yielding structured finding-to-waveform provenance at the original signal level while preserving the distinction between generated interpretation and its supporting evidence.

3.5 Learning Objectives and Optimization

Stage I: Event Grounding. The ECG encoder, CED, and CET are jointly trained with the unbalanced transport objective $\mathcal{L}_{\mathrm{UOT}}$, while BioLORD remains frozen. After convergence, all Stage-I components are frozen to preserve waveform supports and semantic keys.

Stage II: Relational Reasoning and Report Generation. We jointly optimize the objective

$$\min_{\theta} \mathcal{L}_{\mathrm{report}} = \frac{\sum_t -\log p_\theta(y_t \mid y_{<t}, M) + \sum_f w_f \, H(Q_f, P_{\mathrm{comp},f})}{N_{\mathrm{token}} + N_{\mathrm{citation}}},$$

Table 1: ECG report generation performance on MIMIC-IV-ECG across linguistic quality, clinical content, and clinical error metrics. Best and second-best results are highlighted in red and blue, respectively.

Model	Linguistic Quality					Clinical Content			Clinical Error	
	B-1↑	B-4↑	M↑	R-L↑	B-F1↑	F-Pre↑	F-Rec↑	F-F1↑	Hall.↓	Omit.↓
LLaVA-v1.6 Liu et al. (2024b)	0.013	0.000	0.042	0.024	0.775	0.209	0.002	0.004	0.373	0.998
LLaVA-Med Li et al. (2023)	0.025	0.001	0.056	0.044	0.823	0.131	0.004	0.008	0.255	0.996
QOQ-Med Dai et al. (2025)	0.019	0.000	0.074	0.031	0.803	0.441	0.135	0.207	0.372	0.865
MedVLM-R1 Pan et al. (2025)	0.032	0.000	0.113	0.061	0.829	0.357	0.211	0.265	0.322	0.789
PULSE Liu et al. (2026)	0.092	0.018	0.284	0.154	0.847	0.597	0.649	0.622	0.071	0.351
GEM Lan et al. (2025)	0.064	0.004	0.150	0.090	0.841	0.443	0.367	0.401	0.146	0.633
Q-HEART Pham et al. (2025)	0.246	0.165	0.328	0.193	0.845	0.663	0.056	0.103	0.195	0.944
ECG-Chat Zhao et al. (2025)	0.364	0.127	0.316	0.281	0.834	0.698	0.674	0.686	0.394	0.326
ECG-R1 Jin et al. (2026)	0.201	0.122	0.326	0.271	0.852	0.681	0.753	0.715	0.136	0.247
CAMEL Velingker et al. (2026)	0.261	0.109	0.253	0.237	0.814	0.538	0.574	0.555	0.218	0.426
anyECG-chat Li et al. (2026)	0.534	0.436	0.592	0.591	0.847	0.544	0.594	0.568	0.294	0.406
HeartLLM Yang et al. (2026)	0.425	0.281	0.433	0.517	0.858	0.663	0.556	0.605	0.195	0.444
ECG-EGR	**0.876**	**0.734**	**0.871**	**0.843**	**0.962**	**0.849**	**0.829**	**0.839**	**0.045**	**0.171**

where $H(Q_f, P_{\text{comp},f}) = -\sum_{i=1}^{N(X)} Q_{fi} \log P_{\text{comp}}(i \mid f)$. With all Stage-I components frozen, REC, the SetMemory Adapter, Dynamic Event Retriever, and language-model adaptation parameters are trained jointly. The first term supervises report generation, while the second aligns dynamically retrieved Events with the soft Stage-I finding–Event correspondences. Confidence-derived weights w_f emphasize reliable grounding targets without introducing an additional loss-balancing coefficient. Freezing Stage I preserves the learned waveform supports and semantic identities.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. ECG-EGR is trained exclusively on MIMIC-IV-ECG Gow et al. (2023), which provides large-scale 12-lead ECG recordings paired with ECG interpretation text. We partition subjects into 80%, 10%, and 10% training, validation, and test sets using a deterministic subject-level split. All recordings from the same patient are assigned to a single partition, preventing patient overlap and potential information leakage. In addition to in-domain evaluation on MIMIC-IV-ECG, we use PTB-XL Wagner et al. (2020), CPSC2018 Liu et al. (2018), and Chapman-Shaoxing-Ningbo (CSN) Zheng et al. (2020) as external datasets for generalization and cross-task representation transfer. All comparison models are trained using the same training data to ensure a consistent evaluation setting. Detailed dataset statistics and split protocols are provided in Appendix A.1.

Implementation Details. ECG-EGR is optimized in two stages using AdamW with cosine learning-rate scheduling, 5% warmup, gradient clipping, and bf16 mixed precision. Stage I trains the ECG encoder, CED, and CET for 30 epochs to learn clinically grounded Event representations with a learning rate of 2×10^{-4}, while the BioLORD text encoder Remy et al. (2022) remains frozen. Stage II freezes all Stage-I components and trains REC, the SetMemory Adapter, Dynamic Event Retriever, structural-token embeddings, and LoRA parameters of Qwen3-8B for 10 epochs. We use component-specific learning rates of 2×10^{-4}, 1×10^{-4}, 1×10^{-4}, and 2×10^{-5} for REC, memory adaptation, retrieval, and LoRA, respectively. Weight decay is set to 0.01, and the random seed is fixed to 42. Complete optimization and runtime settings are provided in Appendix A.2.

4.2 Evaluation Tasks and Metrics

Report Generation Evaluation. We evaluate ECG-EGR on MIMIC-IV-ECG and PTB-XL to assess in-domain report generation performance and cross-domain generalization across different clinical settings. Linguistic quality is measured using BLEU-1 (B-1), BLEU-4 (B-4), METEOR (M), ROUGE-L (R-L), and BERTScore-F1 (B-F1). Clinical content is assessed through Finding Precision (F-Pre), Finding Recall (F-Rec), and Finding F1 (F-F1), while clinical errors are quantified by Hallucination Rate (Hall.) and Omission Rate (Omit.). Detailed metric definitions and two Event-grounded evidence faithfulness metrics, EG-Nec and EG-Suf, are provided in Appendix A.3.

Table 2: Linear-probing performance on four PTB-XL classification tasks under 1%, 10%, and 100% labeled-data regimes, measured by macro-AUROC (%) to assess representation transfer.

Model	PTB-XL Sub-class			PTB-XL Super-class			PTB-XL Form			PTB-XL Rhythm		
	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
Human-CLAP Takano et al. (2025)	58.06	66.90	74.09	62.83	69.36	72.92	53.41	59.73	67.12	59.98	72.81	80.91
ECGFM-KED Tian et al. (2024)	67.00	76.63	86.57	77.14	82.78	87.37	60.81	69.48	71.54	67.80	80.41	89.60
GEM Lan et al. (2025)	64.94	73.38	84.88	72.47	82.45	86.37	57.85	61.81	74.91	60.94	72.98	84.94
MERL Liu et al. (2024a)	63.95	73.96	80.19	69.98	75.87	81.54	54.63	71.66	79.69	57.89	81.96	87.50
MELP Wang et al. (2025)	79.03	83.78	87.64	76.32	83.76	87.64	63.26	74.95	82.81	85.49	**93.51**	94.26
ECG-Chat Zhao et al. (2025)	68.16	79.37	88.93	75.88	84.77	89.33	59.12	66.73	78.54	67.72	82.29	90.10
HeartLang Jin et al. (2025)	63.92	74.48	84.08	72.06	79.84	85.02	57.54	63.67	76.31	61.57	74.65	85.06
SuPreME Cai et al. (2025)	56.96	63.12	69.14	60.61	63.66	70.47	51.12	55.32	57.28	54.35	64.20	71.73
ST-MEM Na et al. (2024)	64.04	74.54	86.42	71.29	81.90	86.34	59.45	68.52	80.07	68.31	82.36	90.74
Q-HEART Pham et al. (2025)	70.60	83.72	90.10	81.12	88.01	88.87	65.41	75.00	80.04	66.21	80.06	87.02
ECG-EGR	**85.97**	**88.33**	**92.80**	**86.78**	**88.45**	**89.92**	**67.25**	**75.82**	**84.13**	**89.57**	93.46	**96.88**

Cross-Task Representation Transfer. We assess the generalizability of learned Event representations through linear probing on six ECG classification tasks. The pretrained representation is frozen, and only a linear classifier is trained using 1%, 10%, or 100% of the labeled training data. Performance is measured by macro-averaged AUROC. Additional evaluations of Event stability and zero-shot clinical concept transfer are presented in Appendices B.2 and B.3, respectively.

5 Results

5.1 In-Domain ECG Report Generation

Table 1 presents the report generation results on MIMIC-IV-ECG. ECG-EGR achieves the best performance across all ten metrics, demonstrating strong linguistic quality and clinical content consistency. Compared with the second-best result on each metric, ECG-EGR improves BLEU-1 from 0.534 to 0.876, BLEU-4 from 0.436 to 0.734, and BERTScore-F1 from 0.858 to 0.962. Beyond linguistic similarity, ECG-EGR achieves a Finding F1 of 0.839, exceeding ECG-R1 by 0.124, with improvements in both Finding Precision and Finding Recall. This indicates that the generated reports recover a larger fraction of reference findings while maintaining higher precision. Moreover, ECG-EGR reduces the Hallucination Rate to 0.045 and the Omission Rate to 0.171, indicating fewer unsupported clinical findings and missed reference findings. These results are consistent with the Event-grounded formulation, in which waveform-supported Events provide structured evidence for relational composition and subsequent report generation. By organizing ECG information into explicit evidence units rather than relying solely on distributed representations, ECG-EGR provides a structured intermediate representation for generating clinically meaningful findings. Additional out-of-domain report generation results on PTB-XL are provided in Appendix B.1.

5.2 Cross-Task Representation Transfer

Table 2 reports linear-probing performance across four PTB-XL classification tasks under different labeled-data regimes. ECG-EGR achieves the highest AUROC in 11 of 12 settings, demonstrating strong transferability of the learned Event representations beyond report generation. The gains are particularly pronounced under limited supervision. With only 1% labeled data, ECG-EGR outperforms the strongest competing results by 6.94 and 5.66 points on Subclass and Superclass, respectively, indicating that the learned Events retain clinically discriminative information accessible with minimal task-specific supervision. On Rhythm, ECG-EGR reaches 89.57 AUROC with 1% labels, exceeding MELP by 4.08 points. These improvements suggest that support-constrained Event representations capture diagnostic patterns across heterogeneous temporal and lead-wise structures. With full supervision, ECG-EGR remains consistently strong, reaching 92.80 and 96.88 AUROC on Subclass and Rhythm, respectively. MELP marginally outperforms ECG-EGR only on Rhythm with 10% labels (93.51 vs. 93.46). Overall, the results support waveform-grounded Events as effective transferable ECG representations across diverse downstream tasks. Additional cross-dataset transfer results on CPSC2018 and CSN are presented in Appendix B.2.

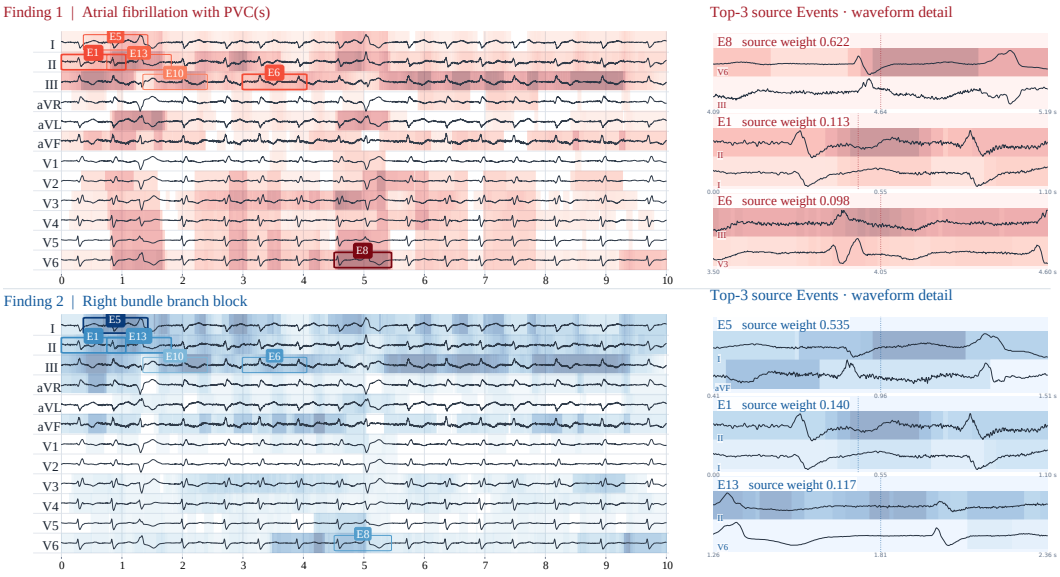


Figure 3: Finding-to-waveform grounding for two generated findings from the same MIMIC-IV-ECG case (ID: 41025048). The upper (red) and lower (blue) rows show finding-specific source-support overlays on the complete 12-lead ECG. All six active Events are labeled, with boxes marking representative high-support cores; the right panels show the three highest-weighted source Events for each finding. Color intensities are normalized separately within each row.

5.3 Qualitative Analysis

Finding-to-Waveform Grounding Case Study.

Figure 3 illustrates finding-specific evidence grounding for a MIMIC-IV-ECG case. The two generated findings describe atrial fibrillation with PVC(s) and right bundle branch block (RBBB). Although both findings share the same inventory of six active Events, they retrieve distinct source-Event distributions. The first finding assigns its highest source weight to Event 8 (0.622), whereas the RBBB finding primarily retrieves Event 5 (0.535). Their top-three Event sets overlap only at Event 1 and account for 83.3% and 79.2% of their source weights. The left panels project each finding's source-weighted Event supports onto the complete 12-lead ECG, while the right panels display waveform details for its three most heavily weighted Events. This visualization makes the finding-conditioned selection of shared evidence units explicit and links generated findings to inspectable waveform regions through the model's structured routing provenance.

Event Relation Graph Visualization. Figure 4 visualizes the REC relation graph for a MIMIC-IV-ECG case with sinus rhythm, premature ventricular contractions, and borderline first-degree atrioventricular block. Nodes represent eight active Events, and directed edges depict pairwise message routing, with weights annotated along each edge. For clarity, low-weight edges are omitted, while the displayed connections reveal non-uniform routing strengths across Event pairs during relational composition. The generated report captures the principal rhythm and conduction findings in the reference interpretation. Removing pairwise relations increases the fixed-report NLL by 0.046, indicating that relational interactions contribute to the likelihood assigned to the generated report. Together, the visualization and intervention illustrate how REC selectively integrates information across waveform-grounded Events.

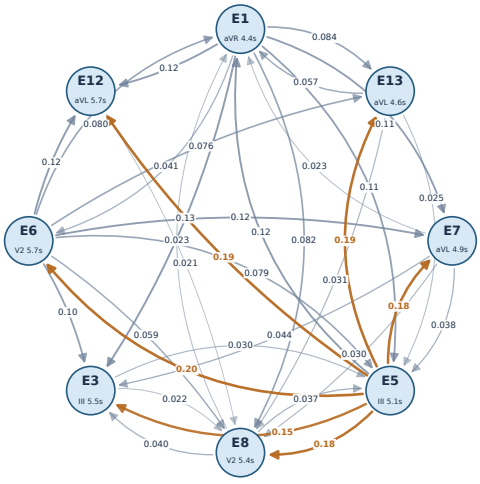


Figure 4: REC relation graph for a MIMIC-IV-ECG case (ID: 48652787).

5.4 Ablation Studies

Representation Ablation. Table 3 compares three frozen representations: Global-Local and Global-Context are obtained by pooling local waveform and contextualized features, respectively, while Event pools waveform-grounded Event representations. Global-Context generally outperforms Global-Local, demonstrating the benefit of contextual information. Event further improves performance across all tasks and annotation budgets, with pronounced gains under limited supervision. At 1% labels, Event exceeds Global-Context by 13.51 and 10.74 AUROC points on PTB-XL Subclass and Rhythm, respectively. These results suggest that organizing contextual information into waveform-grounded Events yields more discriminative and transferable representations than globally pooled local or contextual features. Further details are provided in Appendix B.5.

Table 3: Representation ablation comparing Global-Local, Global-Context, and Event features under 1%, 10%, and 100% labeled-data regimes, measured by macro-AUROC (%). The best results are highlighted in bold.

Dataset	Global-Local			Global-Context			Event		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
PTB-XL Sub-class	70.76	76.81	82.88	72.46	78.11	82.78	**85.97**	**88.33**	**92.80**
PTB-XL Super-class	79.18	82.70	85.47	82.95	85.87	87.69	**86.78**	**88.45**	**89.92**
PTB-XL Form	63.24	66.38	69.67	64.47	70.57	75.85	**67.25**	**75.82**	**84.13**
PTB-XL Rhythm	76.04	76.15	82.13	78.83	82.43	90.25	**89.57**	**93.46**	**96.88**
CPSC2018	69.70	79.01	82.22	82.01	89.03	91.78	**88.85**	**91.65**	**96.04**
Chapman	66.43	71.90	81.05	78.71	85.66	88.45	**82.67**	**88.00**	**97.71**

Component Ablation. Table 4 examines the roles of CED, CET, and REC in report generation and evidence grounding. Replacing CED with learned-query attention produces the largest reductions in EG-Nec and EG-Suf, suggesting that competitive Event discovery helps identify evidence on which generated findings depend. Replacing CET with cosine-softmax alignment degrades clinical precision and grounding scores, indicating that compositional semantic alignment improves the grounding of discovered Events. Removing REC causes the largest degradation in report quality and clinical precision, reducing B-1 from 0.876 to 0.586 and F-Pre from 0.849 to 0.501. Together, these results demonstrate complementary roles for evidence discovery, semantic alignment, and relational composition in generating informative, clinically grounded, and evidence-consistent reports.

Table 4: Ablation study of the proposed components. Event-grounded necessity (EG-Nec) and Event-grounded sufficiency (EG-Suf) are reported to assess evidence faithfulness. Metric definitions are provided in Appendix A.3.

Setting	B-1 ↑	B-F1 ↑	F-Pre ↑	Hall. ↓	EG-Nec ↑	EG-Suf ↑
Full ECG-EGR	**0.876**	**0.962**	**0.849**	**0.045**	**0.094**	**0.131**
CED → Learned-Query Attention	0.612	0.936	0.642	0.112	0.038	0.054
CET → Cosine-Softmax Alignment	0.728	0.944	0.713	0.093	0.052	0.071
w/o REC	0.586	0.924	0.501	0.137	0.061	0.083

6 Conclusion

In this work, we introduced ECG-EGR, an event-grounded framework that connects ECG report generation to localized waveform evidence. Rather than relying solely on global or fixed-window representations, ECG-EGR discovers an adaptive set of Events, each comprising a lead-time support, a stable semantic key, and a composable value. Competitive Event discovery and report-guided semantic alignment establish these evidence units, while relational composition integrates their information without altering their supports or keys. A finding-first generator subsequently produces clinical reports and associates individual findings with Event-derived evidence through semantic retrieval and provenance tracking. Experiments on MIMIC-IV-ECG and out-of-domain PTB-XL demonstrate improvements in linguistic quality, clinical content consistency, and reference-based error metrics. Linear-probing results across multiple ECG classification tasks further support the transferability of the learned Event representations, while component ablations and evidence-faithfulness evaluations clarify the complementary roles of discovery, alignment, and relational composition. Together, these findings establish Event-grounded reasoning as a promising approach to ECG report generation that connects clinical findings with explicit, inspectable waveform evidence.

REFERENCES

Salah S Al-Zaiti, Christian Martin-Gill, Jessica K Zègre-Hemsey, Zeineb Bouzid, Ziad Faramand, Mohammad O Alrawashdeh, Richard E Gregg, Stephanie Helman, Nathan T Riek, Karina Kraevsky-Phillips, et al. Machine learning for ecg diagnosis and risk stratification of occlusion myocardial infarction. *Nature Medicine*, 29(7):1804–1813, 2023.

Mathieu G. G. Bartels, Ivona Najdenkoska, Rutger R van de Leur, Arjan Sammani, Karim Taha, David M Knigge, Pieter A Doevendans, Marcel Worring, and René van Es. Learning to automatically generate accurate ecg captions. In *Proceedings of The 5th International Conference on Medical Imaging with Deep Learning*, volume 172, pp. 86–102, 06–08 Jul 2022.

Amnon Bleich, Antje Linnemann, Bjoern H Diem, and Tim OF Conrad. Automated medical report generation for ecg data: Bridging medical text and signal processing with deep learning. *arXiv preprint arXiv:2412.04067*, 2024.

Mingsheng Cai, Jiuming Jiang, Wenhao Huang, Che Liu, and Rossella Arcucci. SuPreME: A supervised pre-training framework for multimodal ECG representation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11817–11844, 2025.

David Dai, Peilin Chen, Chanakya Ekbote, and Paul Liang. QoQ-Med: Building multimodal clinical foundation models with domain-aware GRPO training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 41978–42025, 2025.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset. *PhysioNet*, 2023. doi: 10.13026/4nqg-sb35. Version 1.0.

William Han, Chaojing Duan, Michael Rosenberg, Emerson Liu, and Ding Zhao. ECG-byte: A tokenizer for end-to-end generative electrocardiogram language modeling. In *Proceedings of the 10th Machine Learning for Healthcare Conference*, volume 298, 15–16 Aug 2025.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. Reading your heart: Learning ecg words and sentences via pre-training ecg language model. In *International Conference on Learning Representations*, volume 2025, pp. 8207–8227, 2025.

Jiarui Jin, Haoyu Wang, Xingliang Wu, Xiaocheng Fang, Xiang Lan, Zihan Wang, Deyun Zhang, Bo Liu, Yingying Zhang, Xian Wu, et al. Ecg-r1: protocol-guided and modality-agnostic mllm for reliable ecg interpretation. *arXiv preprint arXiv:2602.04279*, 2026.

Sunil Vasu Kalmady, Amir Salimi, Weijie Sun, Nariman Sepehrvand, Yousef Nademi, Kevin Bainey, Justin Ezekowitz, Abram Hindle, Finlay McAlister, Russel Greiner, et al. Development and validation of machine learning algorithms based on electrocardiograms for cardiovascular diagnoses at the population level. *npj Digital Medicine*, 7(1):133, 2024.

Akshay Khunte, Veer Sangha, Evangelos K Oikonomou, Lovedeep S Dhingra, Arya Aminorroaya, Andreas Coppi, Sumukh Vasisht Shankar, Elijah Rockers, Bobak J Mortazavi, Deepak L Bhatt, et al. Artificial intelligence-based automated interpretation of images of electrocardiograms: Development and multinational validation of ECG-GPT. *European Heart Journal - Digital Health*, 7(3):ztag031, 2026.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. Clocs: Contrastive learning of cardiac signals across space, time, and patients. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5606–5615, 18–24 Jul 2021.

Jiewei Lai, Huixin Tan, Jinliang Wang, Lei Ji, Jun Guo, Baoshi Han, Yajun Shi, Qianjin Feng, and Wei Yang. Practical intelligent diagnostic algorithm for wearable 12-lead ecg via self-supervised learning on large-scale dataset. *Nature Communications*, 14(1):3741, 2023.

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. In *Advances in Neural Information Processing Systems*, volume 38, pp. 94421–94455, 2025.

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*, volume 36, pp. 28541–28564, 2023.

Haitao Li, Ziyu Li, Yiheng Mao, Ziyi Liu, Zhoujian Sun, and Zhengxing Huang. anyECG-chat: A generalist ECG-MLLM for flexible ECG input and multi-task understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 597–605, 2026.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. Zero-shot ECG classification with multimodal learning and test-time clinical knowledge enhancement. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 31949–31963, 21–27 Jul 2024a.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024b.

Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. Teaching multimodal llms to comprehend 12-lead electrocardiographic images. *npj Digital Medicine*, 9(1):349, 2026.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: An open electrocardiogram foundation model. *JAMIA Open*, 8(5):ooaf122, 2025.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. In *International Conference on Learning Representations*, volume 2024, pp. 15012–15035, 2024.

Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Edward Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 338–353, 2022.

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 (VLMs) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 15966 of *Lecture Notes in Computer Science*, pp. 337–347, 2025.

Hung Manh Pham, Jialu Tang, Aaqib Saeed, and Dong Ma. Q-HEART: ECG question answering via knowledge-informed multimodal LLMs. In *Proceedings of the European Conference on Artificial Intelligence*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 4545–4552, 2025.

Manh Pham Hung, Aaqib Saeed, and Dong Ma. Boosting masked ECG-text auto-encoders as discriminative learners. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 49277–49291, 13–19 Jul 2025.

Yupeng Qiang, Xunde Dong, Xiuling Liu, Fei Hu, and Rongjia Wang. BiECG-LLM: An approach to ECG classification and report generation using a fine-tuned LLM with bi-modal ECG. In *2025 IEEE International Conference on Bioinformatics and Biomedicine*, pp. 7476–7483, 2025.

Jielin Qiu, William Han, Jiacheng Zhu, Mengdi Xu, Michael Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. Transfer knowledge from natural language to electrocardiography: Can we detect cardiovascular disease through language models? In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 442–453, 2023.

François Remy, Kris Demuynck, and Thomas Demeester. BioLORD: Learning ontological representations from definitions for biomedical concepts and their textual descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1454–1465, 2022.

Taisei Takano, Yuki Okamoto, Yusuke Kanamori, Yuki Saito, Ryotaro Nagase, and Hiroshi Saruwatari. Human-CLAP: Human-perception-based contrastive language-audio pretraining. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 131–136, 2025.

Jialu Tang, Tong Xia, Yuan Lu, Cecilia Mascolo, and Aaqib Saeed. Electrocardiogram report generation and question answering via retrieval-augmented self-supervised modeling. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025.

Yuanyuan Tian, Zhiyuan Li, Yanrui Jin, Mengxiao Wang, Xiaoyang Wei, Liqun Zhao, Yunqing Liu, Jinlei Liu, and Chengliang Liu. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. *Cell Reports Medicine*, 5(12):101875, 2024.

Akhil Vaid, Joy Jiang, Ashwin Sawant, Stamatios Lerakis, Edgar Argulian, Yuri Ahuja, Joshua Lampert, Alexander Charney, Hayit Greenspan, Jagat Narula, et al. A foundational vision transformer improves diagnostic performance for electrocardiograms. *npj Digital Medicine*, 6(1):108, 2023.

Neelay Velingker, Alaia Solko-Breslin, Mayank Keoliya, Seewon Choi, Jiayi Xin, Anika Marathe, Alireza Oraii, Rajat Deo, Sameed Khatana, Rajeev Alur, Mayur Naik, and Eric Wong. CAMEL: An ECG language model for forecasting cardiac events. *arXiv preprint arXiv:2602.15677*, 2026.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020.

Zhongwei Wan, Che Liu, Xin Wang, Chaofan Tao, Hui Shen, Jing Xiong, Rossella Arcucci, Huaxiu Yao, and Mi Zhang. MEIT: Multimodal electrocardiogram instruction tuning on large language models for report generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14510–14527, 2025.

Fuying Wang, Jiacheng Xu, and Lequan Yu. From token to rhythm: A multi-scale approach for ECG-language pretraining. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 65059–65074, 13–19 Jul 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jinning Yang, Wenjie Sun, and Wen Shi. HeartLLM: Discretized ECG tokenization for LLM-based diagnostic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34250–34258, 2026.

Kai Yang, Massimo Hong, Jiahuan Zhang, Yizhen Luo, Suyuan Zhao, Ou Zhang, Xiaomao Yu, Jiawen Zhou, Liuqing Yang, Ping Zhang, et al. ECG-LM: Understanding electrocardiogram with a large language model. *Health Data Science*, 5:0221, 2025b.

Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. ECG-Chat: A large ECG-language model for cardiac disease diagnosis. In *2025 IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2025.

Jianwei Zheng, Huimin Chu, Daniele Struppa, Jianming Zhang, Sir Magdi Yacoub, Hesham El-Askary, Anthony Chang, Louis Ehwerhemuepha, Islam Abudayyeh, Alexander Barrett, et al. Optimal multi-stage arrhythmia classification approach. *Scientific Reports*, 10(1):2898, 2020.

Xue Zhou, Tianhui Li, Hiromasa Hayama, Keijiro Nakamura, Shing-Hong Liu, Wenxi Chen, and Xin Zhu. Diagnosis of cardiac conditions from 12-lead electrocardiogram through natural language supervision. *npj Digital Medicine*, 8(1):697, 2025.

AI USE STATEMENT

In this work, we used generative AI tools to assist with scientific writing and language editing, including improving clarity, concision, organization, and presentation of the manuscript. Generative AI was also used to support the interpretation and consistency checking of experimental results, as well as to review the alignment between the described methodology and its implementation. In particular, AI assistance was used to identify potential inconsistencies in terminology, equations, experimental descriptions, reported metrics, and implementation details, and to suggest revisions to improve the precision and readability of the paper. The research methodology, model implementation, experimental execution, and final scientific conclusions were determined by the authors.

All AI-assisted content was reviewed and verified by the authors before inclusion in the manuscript. Scientific claims edited or suggested by generative AI were manually checked against the experimental results, implementation, and relevant literature. Method descriptions were compared with the corresponding code and training or evaluation procedures to ensure consistency with the implemented system. Reported numerical results, comparisons, and derived quantities were verified against the original experimental outputs or recomputed where appropriate. AI-generated suggestions were not accepted without author review, and the authors take full responsibility for the final text, claims, results, and conclusions presented in this work.

ETHICS STATEMENT

This study uses previously released de-identified ECG datasets that are publicly available or accessible under controlled data-use agreements. All datasets are used in accordance with their respective access requirements, licenses, and usage policies. The study does not involve direct interaction with patients, and no attempt is made to re-identify individuals from the available data.

ECG-EGR is developed and evaluated as a research system and is not validated for autonomous clinical decision making or direct deployment in patient care. Generated reports and Event-level evidence may contain errors, omissions, or dataset-specific biases, and performance may vary across institutions, populations, acquisition devices, and reporting practices. Accordingly, the outputs of ECG-EGR should not be interpreted as a substitute for qualified clinical assessment, and its structured provenance should be viewed as model-internal evidence tracing rather than definitive causal or clinical explanation.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training procedure, data splits, optimization settings, evaluation protocols, and metric definitions in the main paper and appendix. The appendix further documents dataset-specific preprocessing, Stage-I and Stage-II training configurations, linear-probing settings, perturbation protocols, zero-shot evaluation procedures, and ablation designs to support reproducibility of the reported results. We also specify the random seed, optimization hyperparameters, and hardware configuration used in our experiments where applicable.

To further facilitate reproducibility, we plan to release the implementation, configuration files, and evaluation scripts upon acceptance of the paper. The released code will include the components required to reproduce the main training and evaluation pipelines described in the manuscript, subject to the access and licensing requirements of the underlying ECG datasets.

A Details of Experimental Settings

A.1 Dataset Details

MIMIC-IV-ECG. MIMIC-IV-ECG Gow et al. (2023) contains approximately 800,000 diagnostic 12-lead ECG recordings from nearly 160,000 patients. Each recording is 10 seconds long and sampled at 500 Hz. We use MIMIC-IV-ECG as the sole training source for ECG-EGR and deterministically partition subjects into 80% training, 10% validation, and 10% test sets. All recordings from the same patient are assigned to a single partition, preventing patient overlap across splits. Its large scale and paired interpretation text make MIMIC-IV-ECG suitable for learning waveform-grounded representations and report-generation objectives. Report-generation evaluation is conducted on ECG–report pairs from the held-out test partition after applying a consistent preprocessing pipeline to normalize report formatting and remove non-informative artifacts. This protocol ensures that model development and final evaluation remain separated at the patient level.

PTB-XL. PTB-XL Wagner et al. (2020) contains 21,837 clinical 12-lead ECG recordings from 18,885 patients, each 10 seconds long and available at 500 Hz. We use PTB-XL for two complementary out-of-domain evaluations. For report generation, we follow ECG-Chat Zhao et al. (2025) and use English-translated free-text reports derived from the original multilingual reports as references. For cross-task representation transfer, we evaluate four PTB-XL classification tasks: diagnostic superclass, diagnostic subclass, form, and rhythm, containing 5, 23, 19, and 12 classes, respectively. We follow the recommended patient-stratified folds, using folds 1–8 for training, fold 9 for validation, and fold 10 for testing.

CPSC2018. CPSC2018 Liu et al. (2018) provides 6,877 public 12-lead ECG recordings sampled at 500 Hz, with recording durations ranging from 6 to 60 seconds. The dataset contains nine diagnostic categories and is used exclusively for external representation-transfer evaluation. Under our evaluation protocol, the `validation_set` in our benchmark copy is reserved for testing, while the `training_set` is deterministically divided into training and validation subsets with a 90/10 ratio. Because the metadata used in our pipeline does not expose patient identifiers, disjointness is enforced and verified at the recording-group level rather than claimed at the patient level.

Chapman-Shaoxing-Ningbo. The Chapman-Shaoxing-Ningbo dataset Zheng et al. (2020) contains 45,152 12-lead ECG recordings collected for large-scale arrhythmia analysis. Each recording contains 5,000 samples per lead at 500 Hz, corresponding to a 10-second ECG. Diagnostic annotations are provided using SNOMED-CT codes and mapped to the benchmark diagnostic categories used for cross-task representation evaluation. We use this dataset only as an external domain and never update the ECG-EGR representation encoder on its labels. Dataset partitions follow the provided stratified folds when available, with group-level disjoint splitting used otherwise.

CODE-15% Dataset. CODE-15% **?** is a large-scale, publicly available 12-lead ECG dataset derived from the Clinical Outcomes in Digital Electrocardiography (CODE) cohort. It contains 345,779 ECG recordings from 233,770 patients, collected by the Telehealth Network of Minas Gerais, Brazil, between 2010 and 2016. The recordings were sampled at 400 Hz and standardized to 4,096 samples per lead through zero-padding. The dataset provides diagnostic annotations for six cardiac abnormalities, including first-degree atrioventricular block, right bundle branch block, left bundle branch block, sinus bradycardia, sinus tachycardia, and atrial fibrillation. In this work, we use CODE-15% to evaluate zero-shot clinical concept transfer with Event Keys, examining whether the learned Event representations can capture clinically meaningful concepts in ECG recordings from an independent cohort.

A.2 Hyperparameter Settings

ECG-EGR follows the two-stage optimization described in Section 3.5. Stage I learns waveform-grounded Event representations by optimizing the ECG encoder, CED, and CET while keeping BioLORD frozen. Stage II loads the frozen Stage-I representation and optimizes relational reasoning and report generation using REC, the SetMemory Adapter, Dynamic Event Retriever, and LoRA adaptation of Qwen3-8B. Both stages use AdamW, cosine learning-rate scheduling, a 5% warmup ratio, gradient clipping at 1.0, and bf16 mixed precision. Both stages are trained on eight

Table 5: Hyperparameter settings for Stage-I Event grounding.

Hyperparameter	Value
Trainable components	ECG encoder, CED, CET
Text encoder	BioLORD-2023 (frozen)
Number of GPUs	8 × RTX 4090
Training epochs	30
Batch size per GPU	8
Gradient accumulation steps	1
Effective batch size	64
Evaluation batch size	8
Learning rate	2×10^{-4}
Optimizer	AdamW
Learning-rate scheduler	Cosine
Warmup ratio	0.05
Weight decay	0.01
Gradient clipping	1.0
Precision	bf16
Number of workers	8

Table 6: Hyperparameter settings for Stage-II relational reasoning and report generation.

Hyperparameter	Value
Language model	Qwen3-8B
Trainable components	REC, memory, retriever, LoRA
Stage-I components	Frozen
Number of GPUs	8 × RTX 4090
Training epochs	10
Batch size per GPU	2
Gradient accumulation steps	2
Effective batch size	32
Evaluation batch size	1
Maximum sequence length	1024
REC learning rate	2×10^{-4}
Memory learning rate	1×10^{-4}
Retriever learning rate	1×10^{-4}
LoRA learning rate	2×10^{-5}
Optimizer	AdamW
Learning-rate scheduler	Cosine
Warmup ratio	0.05
Weight decay	0.01
Gradient clipping	1.0
Precision	bf16
Gradient checkpointing	Enabled
Attention implementation	SDPA
Number of workers	4

NVIDIA GeForce RTX 4090 GPUs, with effective batch sizes of 64 and 32 for Stage I and Stage II, respectively. Detailed hyperparameters are summarized in Tables 5 and 6.

A.3 Evaluation Metrics

Linguistic Quality. We evaluate the linguistic similarity between generated and reference reports using BLEU-1, BLEU-4, METEOR, ROUGE-L, and BERTScore-F1. BLEU measures n-gram precision, METEOR accounts for lexical correspondence, ROUGE-L evaluates longest-common-subsequence overlap, and BERTScore-F1 measures contextual semantic similarity. Structural tokens introduced by ECG-EGR are removed before evaluation, and all models are assessed using the same text normalization and metric implementations.

Clinical Content. Linguistic similarity alone may not reflect whether a generated report correctly identifies clinically relevant ECG abnormalities. We therefore employ a deterministic ECG finding parser to map generated and reference reports into a shared set of canonical clinical concepts.

The parser normalizes synonymous expressions and distinguishes positive, uncertain, and explicitly negated findings. Positive and uncertain mentions are treated as reported findings, whereas explicitly negated findings are excluded from concept matching. Finding Precision, Finding Recall, and Finding F1 are computed from the resulting concept sets using micro-aggregated true positives, false positives, and false negatives. Specifically, Finding Precision measures the proportion of generated findings present in the reference, while Finding Recall measures the proportion of reference findings recovered by the generated report. Finding F1 is their harmonic mean. This concept-based evaluation is independent of the BioLORD encoder used for Event semantic alignment.

Clinical Error. We additionally report Hallucination Rate and Omission Rate to express clinical disagreement from an error-oriented perspective. Because ECG reports may be incomplete, a generated finding absent from the reference is not automatically considered a hallucination. Instead, Hallucination Rate measures the proportion of generated reported findings explicitly contradicted by the reference, including explicit reference negation, incompatibility with a strictly normal reference, or predefined mutually exclusive clinical concepts. Omission Rate measures the proportion of reference findings not recovered by the generated report. With N_{contra} denoting the number of contradicted generated findings, N_{pred} the total number of generated reported findings, and N_{ref} the total number of reference reported findings, the two metrics are defined as

$$\text{Hallucination Rate} = \frac{N_{\mathrm{contra}}}{N_{\mathrm{pred}}}, \qquad \text{Omission Rate} = \frac{\mathrm{FN}}{N_{\mathrm{ref}}}.$$

These metrics quantify disagreement with the available reference rather than independently establishing the clinical correctness of every generated statement.

Evidence Faithfulness. To assess whether retrieved Events contribute to the generated findings, we introduce Event-Grounded Necessity (EG-Nec) and Event-Grounded Sufficiency (EG-Suf). Unlike report-level quality metrics, these measures evaluate the prediction dependence of ECG-EGR on its retrieved evidence through controlled Event interventions. For a generated finding f, let $\mathcal{E}$ denote the set of active composed Events and $\mathcal{C}_f$ its retrieved Event subset. We select $\mathcal{C}_f$ from the finding-conditioned retrieval distribution using a cumulative probability threshold of 0.8, retaining between one and four Events. We additionally construct a collection $\mathcal{R}_f$ of control Event subsets matched to $\mathcal{C}_f$ in cardinality and approximately matched in waveform-support mass.

Let $\ell_f(\mathcal{A})$ denote the mean token-level negative log-likelihood of the fixed generated finding f under teacher forcing when only the Events in $\mathcal{A}$ are available to the decoder. EG-Nec compares the effect of removing retrieved Events against removing matched control Events:

$$\text{EG-Nec}_f = \ell_f(\mathcal{E} \setminus \mathcal{C}_f) - \frac{1}{|\mathcal{R}_f|} \sum_{\mathcal{R} \in \mathcal{R}_f} \ell_f(\mathcal{E} \setminus \mathcal{R}).$$

A positive EG-Nec indicates that removing the retrieved Events increases the finding NLL more than removing comparable control Events, suggesting that the retrieved evidence is relatively necessary for the model's prediction.

EG-Suf evaluates whether the retrieved Events alone provide greater support for the finding than matched control subsets:

$$\text{EG-Suf}_f = \frac{1}{|\mathcal{R}_f|} \sum_{\mathcal{R} \in \mathcal{R}_f} \ell_f(\mathcal{R}) - \ell_f(\mathcal{C}_f).$$

A positive EG-Suf indicates that the retrieved Events preserve the generated finding with lower prediction loss than comparable control evidence. We average both metrics across eligible findings, excluding cases without valid matched controls and, for EG-Nec, cases where removing the retrieved subset would eliminate all active Events. Both metrics are reported with higher values indicating stronger relative evidence dependence. They assess the faithfulness of Event retrieval to the model's own predictions rather than the clinical correctness of those predictions.

B More Results

B.1 Out-of-Domain ECG Report Generation

Table 7 presents report generation results on PTB-XL, evaluating cross-domain generalization without target-domain fine-tuning. ECG-EGR achieves the best performance across all ten metrics,

Table 7: Out-of-domain ECG report generation on PTB-XL without target-domain fine-tuning, evaluated across linguistic quality, clinical content, and clinical error metrics.

Model	Linguistic Quality					Clinical Content			Clinical Error	
	B-1↑	B-4↑	M↑	R-L↑	B-F1↑	F-Pre↑	F-Rec↑	F-F1↑	Hall.↓	Omit.↓
LLaVA-v1.6 Liu et al. (2024b)	0.021	0.001	0.057	0.021	0.765	0.264	0.009	0.017	0.400	0.991
LLaVA-Med Li et al. (2023)	0.051	0.001	0.079	0.049	0.813	0.092	0.004	0.008	0.205	0.996
QOQ-Med Dai et al. (2025)	0.027	0.002	0.090	0.034	0.792	0.420	0.167	0.239	0.325	0.833
MedVLM-R1 Pan et al. (2025)	0.037	0.002	0.111	0.060	0.788	0.681	0.220	0.333	0.104	0.780
PULSE Liu et al. (2026)	0.083	0.008	0.179	0.095	0.822	0.400	0.618	0.486	0.134	0.382
GEM Lan et al. (2025)	0.079	0.001	0.149	0.088	0.829	0.413	0.452	0.432	0.141	0.548
Q-HEART Pham et al. (2025)	0.034	0.001	0.033	0.017	0.743	0.657	0.063	0.115	0.183	0.937
ECG-Chat Zhao et al. (2025)	0.171	0.083	0.154	0.129	0.795	0.215	0.405	0.281	0.225	0.595
ECG-R1 Jin et al. (2026)	0.043	0.006	0.170	0.064	0.809	0.361	0.571	0.442	0.122	0.429
CAMEL Velingker et al. (2026)	0.295	0.140	0.290	0.285	0.826	0.351	0.382	0.366	0.306	0.618
anyECG-chat Li et al. (2026)	0.201	0.112	0.298	0.202	0.820	0.219	0.322	0.261	0.275	0.678
HeartLLM Yang et al. (2026)	0.334	0.105	0.233	0.217	0.843	0.657	0.563	0.606	0.162	0.437
ECG-EGR	**0.412**	**0.239**	**0.468**	**0.471**	**0.861**	**0.761**	**0.666**	**0.710**	**0.046**	**0.334**

demonstrating its ability to transfer ECG interpretation knowledge beyond the training distribution. Compared with the second-best result on each metric, ECG-EGR improves BLEU-1 from 0.334 to 0.412, BLEU-4 from 0.140 to 0.239, and ROUGE-L from 0.285 to 0.471, indicating greater consistency with reference reports despite differences in reporting styles and data distributions. More importantly, ECG-EGR achieves a Finding F1 of 0.710, exceeding HeartLLM by 0.104, with improvements in both Finding Precision and Recall. Its Hallucination Rate of 0.046 and Omission Rate of 0.334 further indicate fewer reference-contradicted findings and missed reference findings. Although performance on most metrics decreases relative to the in-domain MIMIC-IV-ECG evaluation, ECG-EGR maintains consistent advantages across linguistic and clinical metrics. These findings suggest that organizing ECG signals into waveform-grounded Events provides transferable clinical information for report generation, while relational composition enables integration of evidence across temporal and lead-specific patterns. Overall, the results support the generalizability of Event-grounded ECG interpretation across distinct clinical datasets.

B.2 Cross-Dataset Representation Transfer

Table 8: Cross-dataset linear-probing performance on CPSC2018 and CSN under 1%, 10%, and 100% labeled-data regimes, measured by macro-AUROC (%).

Model	CPSC2018			CSN		
	1%	10%	100%	1%	10%	100%
Human-CLAP Takano et al. (2025)	65.53	74.81	84.92	63.57	75.29	85.51
ECGFM-KED Tian et al. (2024)	75.95	88.03	93.83	70.13	81.48	94.69
GEM Lan et al. (2025)	70.06	82.26	94.62	68.14	77.73	89.92
MERL Liu et al. (2024a)	70.83	84.94	89.27	67.33	82.69	87.59
MELP Wang et al. (2025)	81.58	90.85	95.06	74.56	85.91	96.60
ECG-Chat Zhao et al. (2025)	75.72	89.54	95.15	74.59	85.38	95.44
HeartLang Jin et al. (2025)	66.73	81.44	94.61	67.96	78.28	88.85
SuPreME Cai et al. (2025)	56.68	66.74	74.77	58.58	66.57	76.52
ST-MEM Na et al. (2024)	76.18	89.11	95.87	71.56	84.02	95.48
Q-HEART Pham et al. (2025)	76.14	86.84	94.59	71.45	82.46	93.50
ECG-EGR	**88.85**	**91.65**	**96.04**	**82.67**	**88.00**	**97.71**

To further assess the generalizability of ECG-EGR beyond the source domain, we evaluate frozen Stage-I Event representations on CPSC2018 and CSN using the same linear-probing protocol as PTB-XL. As shown in Table 8, ECG-EGR achieves the highest macro-AUROC across all six settings, demonstrating consistent transferability across datasets and label spaces. The gains are particularly pronounced under limited supervision. With only 1% labeled data, ECG-EGR achieves macro-AUROCs of 88.85 on CPSC2018 and 82.67 on CSN, surpassing the strongest baselines by 7.27 and 8.08 percentage points, respectively. These gains suggest that waveform-grounded Events retain clinically discriminative information that can be effectively exploited with minimal target-

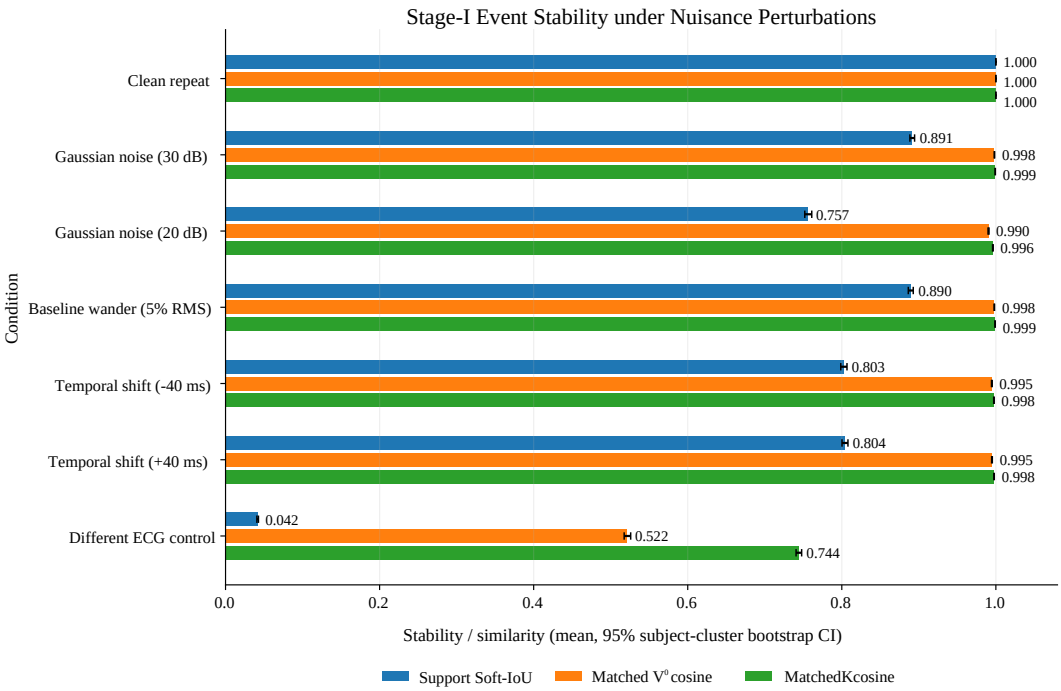


Figure 5: Event stability under nuisance perturbations. Bars report cardinality-penalized support Soft-IoU, matched V^0 cosine similarity, and matched Event Key K cosine similarity, with 95% subject-cluster bootstrap confidence intervals. Clean repeat provides a numerical stability reference, while the different-ECG condition serves as an instance-specific negative control.

domain supervision. As the labeled training fraction increases, ECG-EGR maintains its advantage, reaching 96.04 and 97.71 AUROC on CPSC2018 and CSN with full supervision. However, the performance margin narrows on CPSC2018, indicating that competing representations become increasingly effective as more task-specific labels become available. Together with the PTB-XL results, these findings support the cross-dataset transferability of learned Event representations. By combining context-aware evidence discovery with support-constrained waveform reading, ECG-EGR provides adaptive intermediate representations that preserve diagnostically useful information across heterogeneous ECG classification tasks without updating the pretrained representation.

B.3 Event Stability under Nuisance Perturbations

We examine whether the frozen Stage-I model recovers consistent sample-specific Events when the same ECG undergoes nuisance perturbations. The evaluation uses 5,000 held-out MIMIC-IV-ECG recordings and considers Gaussian noise at 30 and 20 dB SNR, 0.33-Hz baseline wander with an amplitude of 5% of per-lead RMS, and temporal shifts of ± 40 ms. For temporal shifts, the known displacement is undone on the perturbed supports before matching, thereby evaluating equivariance rather than coordinate agreement. Active Events from clean and perturbed signals are matched by Hungarian assignment using only support Soft-IoU, without using Event indices, values, or semantic keys. We report cardinality-penalized support Soft-IoU, cosine similarity of matched initial Event states V^0, and cosine similarity of matched Event Keys K. Clean repeat and different-ECG conditions provide numerical-stability and instance-specific controls, respectively.

Figure 5 shows that Event representations remain more consistent under nuisance perturbations than across different ECGs. Under 30 dB noise and baseline wander, support Soft-IoU remains above 0.91, while matched V^0 cosine similarity remains around 0.89. Stronger 20 dB noise reduces these values to 0.812 and 0.757, while temporal shifts yield support similarity around 0.85. In contrast, Event Keys remain highly stable across perturbations, with matched cosine similarity between 0.990 and 0.998. The different-ECG control drops sharply to 0.058, 0.042, and 0.522 for support, V^0, and K, respectively. This separation supports the perturbation stability and instance specificity of learned Events while allowing support and evidence states to respond to signal perturbations.

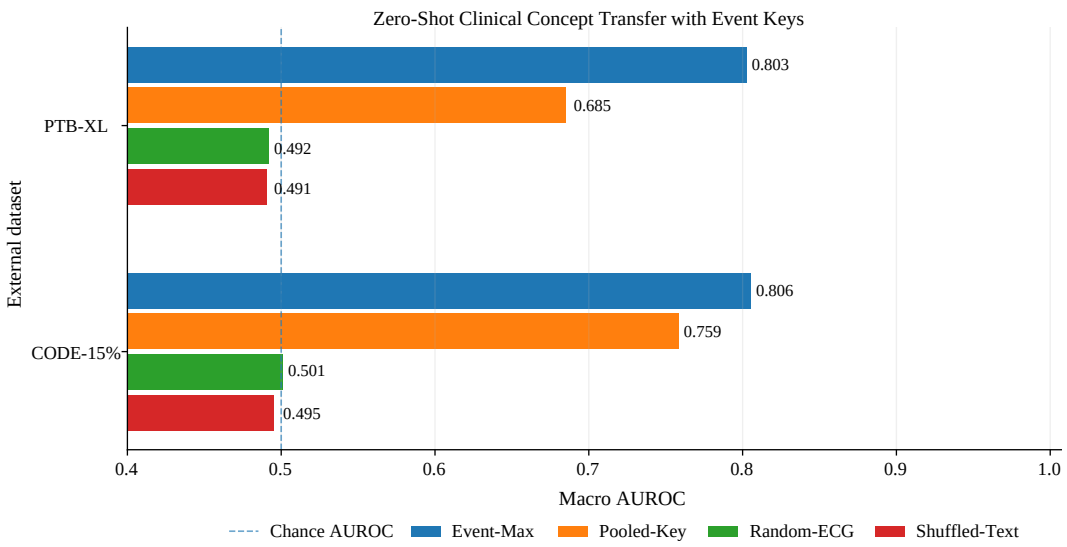


Figure 6: Zero-shot clinical concept transfer with frozen Event Keys on PTB-XL and CODE-15%. Macro-AUROC is reported for Event-Max, Pooled-Key, Random-ECG, and Shuffled-Text, with 95% patient/group-cluster bootstrap confidence intervals for Event-Max and Pooled-Key.

B.4 Zero-Shot Clinical Concept Transfer with Event Keys

We further examine whether the semantic addresses learned by Stage I transfer to external clinical concepts across distinct datasets and diagnostic label spaces without target-domain adaptation. For each target concept c, a fixed textual description is encoded by the frozen BioLORD encoder to obtain a prototype t_c. Given an ECG, the frozen Stage-I model produces a set of active Event Keys $\{K_i\}$, and the zero-shot score is defined as the maximum cosine similarity between t_c and any active Event Key. No target-domain classifier is trained, and no training or validation labels are used for parameter fitting, calibration, or threshold selection. Target labels are used only to define the evaluation concepts and compute final test metrics. We compare this Event-Max score with a Pooled-Key representation obtained by averaging and normalizing the active Event Keys, together with Random-ECG and Shuffled-Text controls. Performance is evaluated using macro AUROC.

As shown in Figure 6, Event-Max provides the strongest zero-shot discrimination on both PTB-XL and CODE-15%, consistently exceeding the Pooled-Key representation and both negative controls. The improvement over Pooled-Key indicates that preserving localized semantic addresses is more informative than collapsing all Events into a single ECG-level representation. Meanwhile, the substantially weaker Random-ECG and Shuffled-Text results indicate that performance depends on both ECG-specific Event representations and their alignment with the corresponding clinical concepts. These results provide complementary evidence that Event Keys retain transferable clinical semantics beyond the report-generation domain, while not implying that the corresponding concepts were absent from MIMIC-IV-ECG supervision.

B.5 Representation Ablation

Forest-Plot Analysis. Figure 7 presents a forest-plot comparison of three frozen ECG representations across eight classification tasks under 1%, 10%, and 100% labeled-data settings. Global-Local and Global-Context are obtained by globally pooling local waveform features and contextualized features, respectively, whereas Event pools learned waveform-grounded Event representations. All three representations are evaluated using the same linear-probing protocol, allowing their downstream discriminative utility to be compared without updating the pretrained encoder. Global-Context generally improves upon Global-Local, indicating that contextual interaction contributes useful information beyond local waveform features. Event achieves the highest mean AUROC in all 24 settings, although the magnitude of improvement varies across tasks and annotation budgets. At 1% labels, Event exceeds Global-Context by 13.51 AUROC points on PTB-XL Subclass and by 10.74 points on PTB-XL Rhythm, whereas the corresponding gain on SPH is only 0.42 points. Un-

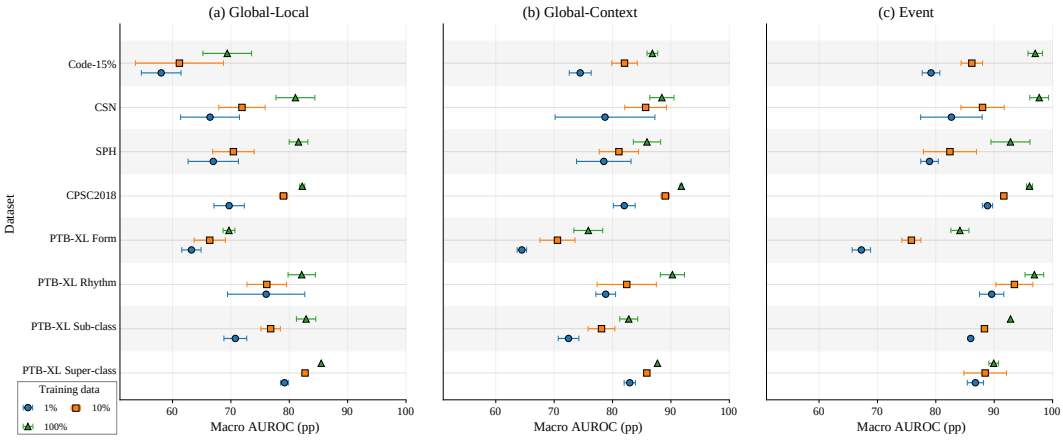


Figure 7: Forest-plot comparison of three ECG representations across eight classification tasks under 1%, 10%, and 100% labeled-data settings. Panels (a), (b), and (c) show Global-Local, Global-Context, and Event representations, respectively. Markers indicate mean macro-AUROC, and error bars denote ± one standard deviation.

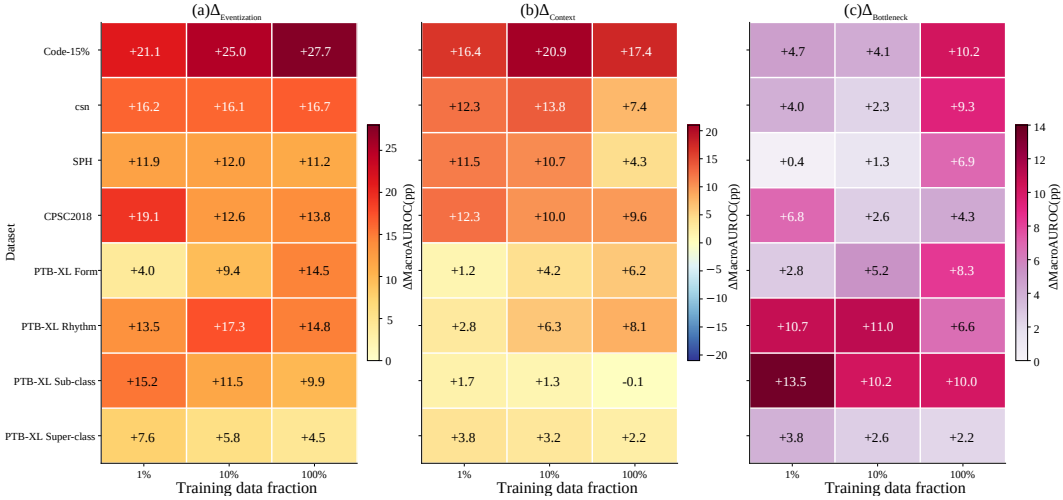


Figure 8: Heatmaps of representation-level macro-AUROC gains across eight ECG classification tasks under 1%, 10%, and 100% labeled-data settings. Panels (a)–(c) show $\Delta_{\text{Eventization}}$, Δ_{Context}, and $\Delta_{\text{Bottleneck}}$, respectively. Cell values indicate mean AUROC differences, with positive values denoting improvements. Each panel uses an independent color scale.

der full supervision, Event reaches 0.9703 AUROC on CODE-15% and 0.9771 on CSN, compared with 0.8682 and 0.8845 for Global-Context. Variability also differs across datasets and settings, so the observed margins should not be interpreted as uniformly precise. Overall, these results suggest that adaptive Event representations retain discriminative information beyond that captured by global pooling, with benefits that depend on the downstream task and available supervision.

Decomposition of Representation Gains. Figure 8 compares three macro-AUROC gains across eight tasks and three annotation budgets under a shared linear-probing protocol with frozen pre-trained features: $\Delta_{\text{Eventization}}$ (Event minus Global-Local), Δ_{Context} (Global-Context minus Global-Local), and $\Delta_{\text{Bottleneck}}$ (Event minus Global-Context). Eventization yields positive gains in all 24 settings, reaching 27.66 AUROC points on CODE-15% with full supervision. Contextualization improves performance in nearly every setting, except PTB-XL Subclass with 100% labels (-0.10 points). Event consistently outperforms Global-Context, with gains of 13.51 points on PTB-XL Subclass at 1% labels and 11.03 points on PTB-XL Rhythm at 10%. These results suggest that adaptive Event aggregation provides additional discriminative information beyond contextualization, although the magnitude of improvement varies across tasks and annotation budgets.