PHYSIOTRACE: TRACING ZERO-SHOT ECG DECISIONS THROUGH CLINICAL OBSERVATIONS TO WAVEFORM EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

ECG-language models enable flexible zero-shot diagnosis, yet their decisions remain difficult to interpret. Existing ECG explanations often focus on signal-level attribution that identifies *where* evidence lies in the waveform, while emerging ECG-language methods begin to expose *what* clinical observations support a diagnosis. However, these two views remain disconnected. We introduce **PhysioTrace**, a hierarchical explanation framework that makes ECG predictions traceable from diagnosis to clinical observation to waveform evidence. Its key idea is to use task-agnostic clinical observations as a shared intermediate representation: each diagnosis is decomposed into observation-level contributions, and each observation is grounded in the multilead ECG through observation-conditioned counterfactual attribution. Experiments across multiple ECG datasets and backbones show that PhysioTrace largely preserves zero-shot performance while providing faithful observation-level explanations and observation-specific waveform grounding. PhysioTrace therefore turns disconnected semantic and signal-level explanations into a unified and auditable reasoning structure. Code is available at `https://anonymous.4open.science/r/PhysioTrace/`.

1 INTRODUCTION

ECG–language models enable flexible zero-shot cardiac diagnosis by matching ECG waveforms with natural-language descriptions (Liu et al., 2024). However, their predictions remain difficult to interpret: the matching score indicates *what* the model predicts, but not the clinical finding or waveform evidence supporting the decision Johnson et al. (2025); Siontis et al. (2021).

Existing explanations provide only partial solutions. Attribution methods identify *where* the model responds in the waveform but do not reveal the corresponding clinical observation. Conversely, concept-based methods such as ZETA (Tang et al., 2026) introduce clinically meaningful observations but lack patient-specific waveform grounding. Thus, semantic explanations and signal-level evidence remain disconnected.

Why this is hard. Connecting these two levels requires a shared representation between semantic decisions and waveform signals. Such a representation must simultaneously satisfy three requirements: it should correspond to clinically recognizable observations, preserve the original zero-shot prediction of the frozen ECG–language model, and provide patient-specific waveform evidence rather than generic descriptions.

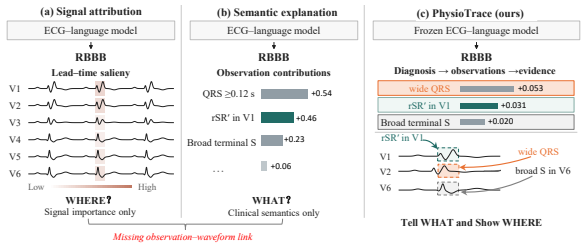

Figure 1: **Motivation.** PhysioTrace connects diagnosis-level observations with patient-specific waveform evidence through a shared observation representation.

Our solution. We introduce **PhysioTrace**, which learns a task-agnostic clinical observation space as the missing interface between diagnosis and waveform evidence:

$$\text{Diagnosis} \rightarrow \text{Clinical Observation} \rightarrow \text{Waveform Evidence}.$$

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.

Given a frozen ECG–language model, PhysioTrace projects ECGs and diagnostic prompts into the shared observation space, enabling each prediction to be decomposed into observation-level contributions. The same observation coordinates are then grounded to lead–time waveform evidence through counterfactual intervention. Therefore, the observation explaining *why* a diagnosis is predicted is also the variable that identifies *where* its supporting evidence appears. Across multiple ECG datasets and ECG–language backbones, PhysioTrace preserves zero-shot performance while providing clinically validated observations, faithful prediction decomposition, and observation-specific waveform grounding.

Our contributions are threefold:

- We introduce a clinically grounded observation space that decomposes zero-shot ECG diagnosis into interpretable observations and their corresponding waveform evidence.
- Unlike diagnosis-specific observation, our observation space is task-agnostic and derived from real ECG reports, enabling explanations that transfer across diagnoses and datasets.
- We establish faithful links between observations, diagnoses, and waveform evidence through recognition, intervention, and grounding evaluations.

2 Related Work

ECG-language Models and Zero-shot. ECG-language models align ECG signals with clinical text to enable flexible zero-shot diagnosis. MERL Liu et al. (2024) established this paradigm through contrastive ECG-report pretraining and text-prompted inference. Subsequent methods further improve ECG-language alignment, including D-BETA Pham et al. (2024), MELP Wang et al. (2025), and SGERA Chen et al., with related efforts exploring generative pretraining and alternative language supervision Zhao et al. (2025); Kwak et al. (2026); Wan et al. (2025). While these models achieve strong zero-shot performance, their decisions remain difficult to interpret.

Explainability for ECG Models Existing ECG explanations mainly highlight important leads or waveform regions using gradient- or perturbation-based attribution (Plagwitz et al., 2024; Wagner et al., 2024; Storås et al., 2025; Suh et al., 2024). Such pointwise maps poorly reflect clinical ECG reasoning, which relies on structured morphology, rhythm, and cross-lead relationships. They show *where* evidence lies, but not *what* clinical observation it represents. More broadly, post-hoc concept bottlenecks use semantic variables to explain pretrained classifiers (Koh et al., 2020; Yang et al., 2023), but do not necessarily connect those variables to patient-specific waveform evidence.

Recent ECG-language methods begin to expose this semantic layer. ZETA (Tang et al., 2026), for example, explains predictions through clinical observations, but does not ground them to patient-specific waveform evidence. PhysioTrace instead organizes diagnosis, observations, and waveform evidence into a unified hierarchy, making predictions systematically traceable and auditable.

3 Method

PhysioTrace explains ECG–language predictions through a shared clinical observation representation. The framework addresses two complementary requirements: diagnostic scores should be decomposable into named clinical observations, and each observation should be traceable to patient-specific waveform evidence. We develop a joint ECG–text observation decomposition for the first requirement and observation-conditioned lead–time interventions for the second. Both components operate through the same observation coordinates, allowing waveform responses to be traced to individual diagnostic contributions. The ECG and text encoders remain frozen throughout; only the observation projection is learned.

3.1 Problem Setup

Let $x \in \mathbb{R}^{L\times T}$ denote a multilead ECG, where $L = 12$ and T is the number of temporal samples. A pretrained ECG–language model maps the ECG and a diagnostic prompt y_k to normalized embeddings

$$v = f_{\text{ecg}}(x) \in \mathbb{R}^d, \qquad t_k = f_{\text{text}}(y_k) \in \mathbb{R}^d. \tag{1}$$

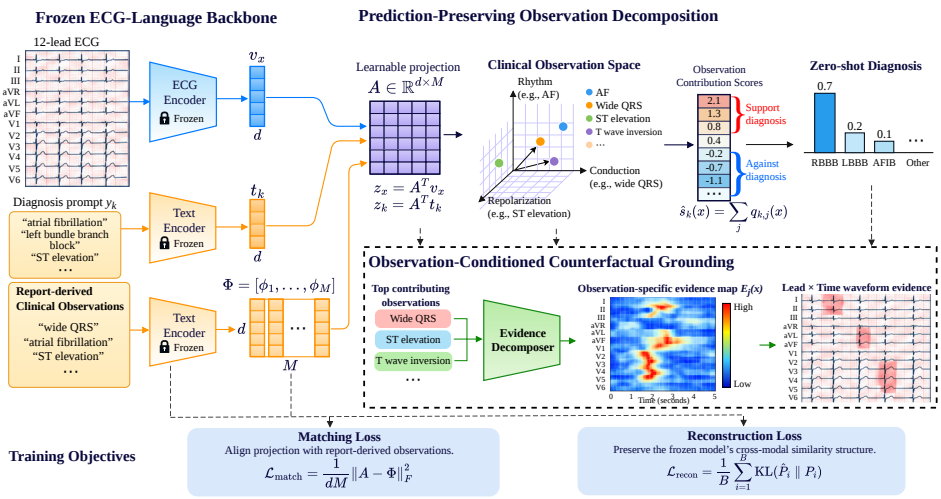

Figure 2: **Overview of PhysioTrace.** PhysioTrace decomposes ECG-language predictions into clinical observation-level contributions and further grounds these observations to lead–time waveform evidence through counterfactual intervention.

Both encoders include their projection heads and L2 normalization. The original zero-shot score for diagnosis k is

$$s_k(x) = v^\top t_k, \tag{2}$$

with backbone-specific logit scaling omitted for clarity.

This global similarity provides neither an explicit decomposition into clinical observations nor an assignment of those observations to lead–time regions. Our objective is to construct an explanation

$$\text{Diagnosis } k \longrightarrow \text{Observation } c_j \longrightarrow \text{Lead–time evidence}, \tag{3}$$

in which the observation used to explain the diagnostic score is also the variable queried during waveform intervention. The observation decomposition defines a projected score $\hat{s}_k(x)$ whose agreement with the original score $s_k(x)$ is encouraged during training and evaluated empirically.

3.2 Report-Derived Clinical Observation Space

We construct a shared observation vocabulary from source-domain ECG reports so that the same representation can be reused across downstream diagnostic tasks. The vocabulary targets findings describable from ECG waveforms, including rhythm, conduction, morphology, and repolarization observations.

We extract and normalize frequent ECG-related phrases from 771,693 MIMIC-IV-ECG reports, obtaining 1,239 candidates after frequency- and quality-based filtering. Curation removes diagnostic names and aliases, report-level expressions, technical descriptions, and other non-waveform terms. Mixed phrases are decomposed to retain their observable components, and redundant normalized phrases are conservatively merged. This process produces a vocabulary $\mathcal{C} = \{c_1, \ldots, c_M\}$ with $M = 351$ observations. Clinical experts assess the retained phrases for their clinical meaning and waveform observability.

The frozen text encoder provides a semantic direction for each observation:

$$\phi_j = \frac{f_{\text{text}}(c_j)}{\|f_{\text{text}}(c_j)\|_2}, \qquad \Phi = [\phi_1, \ldots, \phi_M] \in \mathbb{R}^{d \times M}. \tag{4}$$

These directions anchor the observation coordinates used in the decomposition below.

3.3 Joint ECG–Text Observation Decomposition

The observation decomposition expresses ECG signals and diagnostic prompts in a common clinical coordinate system. Its design must balance two requirements: the coordinates should retain the

semantics of the report-derived observations, and their interaction should approximate the diagnostic behavior of the pretrained model.

Shared observation coordinates. We learn a projection $A \in \mathbb{R}^{d\times M}$ and apply it to both modalities:

$$z_x = A^\top v, \qquad z_k = A^\top t_k. \tag{5}$$

Here, $z_x^{(j)}$ is the ECG response along observation coordinate j, and $z_k^{(j)}$ specifies how that coordinate contributes to the diagnostic prompt. Applying the same projection to both modalities gives

$$\hat{s}_k(x) = z_x^\top z_k = v^\top A A^\top t_k = \sum_{j=1}^{M} q_{k,j}(x), \qquad q_{k,j}(x) = z_x^{(j)} z_k^{(j)}. \tag{6}$$

Thus, Each observation contribution depends jointly on the patient ECG and the diagnosis being explained. Positive and negative values of $q_{k,j}(x)$ respectively support and oppose the projected diagnostic score. The decomposition is exact for $\hat{s}_k(x)$, whereas its agreement with the original score $s_k(x)$ is approximate and evaluated empirically. The observation-contribution and waveform-intervention analyses below therefore characterize the projected scoring function, rather than an exact decomposition of the frozen model's original score.

The projection is shared across diagnostic prompts. A new diagnosis can therefore be evaluated by projecting its text embedding into the existing observation space, without learning a separate diagnostic classifier or reconstructing a class-specific vocabulary.

Learning semantic and predictive structure. We optimize only A using source-domain ECG–report pairs. For a batch of B pairs, let $V = [v_1^\top; \ldots; v_B^\top] \in \mathbb{R}^{B\times d}$ and $U = [u_1^\top; \ldots; u_B^\top] \in \mathbb{R}^{B\times d}$, where u_i is the normalized embedding of the paired report.

A semantic matching term anchors the learned directions to the observation embeddings:

$$\mathcal{L}_{\text{match}} = \frac{1}{dM}\|A - \Phi\|_F^2. \tag{7}$$

To retain the cross-modal relationships of the frozen model, we also match its ECG-to-report similarity distributions:

$$P = \text{Softmax}_{\text{row}}(VU^\top/\tau), \qquad \hat{P} = \text{Softmax}_{\text{row}}\big((VA)(UA)^\top/\tau\big). \tag{8}$$

We use the same temperature $\tau = 1$ for both distributions, without an additional learnable logit scale. The reconstruction term and joint objective are

$$\mathcal{L}_{\text{recon}} = \frac{1}{B}\sum_{i=1}^{B} D_{\text{KL}}(\hat{P}_i \,\|\, P_i), \qquad \mathcal{L}_{\text{obs}} = \mathcal{L}_{\text{match}} + \lambda\mathcal{L}_{\text{recon}}. \tag{9}$$

The two terms respectively encourage interpretable observation directions and preservation of the pretrained cross-modal geometry. Projection training uses no downstream diagnostic labels or waveform localization annotations.

3.4 Observation-Conditioned Lead–Time Grounding

The decomposition identifies which observation coordinates contribute to a diagnosis. We next locate waveform regions that affect each coordinate. Because the observation representation is computed from a global ECG embedding, we obtain localization by querying the model under input interventions. This procedure requires forward evaluations of the ECG encoder and does not require local feature maps, token-level representations, or a separately trained localizer.

Structured waveform interventions. We partition the ECG into regions

$$\mathcal{R} = \{r_{\ell,w} : \ell \in \{1,\ldots,L\},\ w \in \{1,\ldots,W\}\}, \tag{10}$$

where each region contains one lead and one temporal window. For a 10-second ECG, the default setting uses $W = 40$ windows, giving $12 \times 40 = 480$ regions.

For each region r, we construct $x_{\setminus r} = \mathcal{I}(x, r)$ by replacing its samples with linear interpolation between the surrounding boundary values. All samples outside the region remain unchanged. The operator suppresses local morphology while avoiding an abrupt zero mask. It defines a controlled model perturbation and is not assumed to produce a physiologically realizable ECG.

Observation-specific responses. We measure the response of observation j to the intervention as

$$E_j(r) = [A^\top f_{\text{ecg}}(x)]_j - [A^\top f_{\text{ecg}}(x_{\setminus r})]_j. \tag{11}$$

A positive value means that intervening on r reduces the observation coordinate; a negative value means that it increases the coordinate. Evaluating this response across the lattice produces $E_j \in \mathbb{R}^{L\times W}$.

Conditioning on j allows observations associated with the same diagnosis to be queried separately. For example, conduction-related and repolarization-related observations can be assigned different lead–time responses within the same ECG. The resulting maps describe sensitivity of the learned observation coordinates under the specified intervention. Their clinical correspondence and specificity are evaluated separately.

3.5 Tracing Waveform Effects to Diagnostic Contributions

The shared observation representation connects the decomposition and grounding components quantitatively. For a diagnosis k and observation j, the change in diagnostic contribution caused by intervening on region r is

$$G_{k,j}(r) := q_{k,j}(x) - q_{k,j}(x_{\setminus r}) = z_k^{(j)} E_j(r). \tag{12}$$

This identity follows directly from the shared decomposition: the diagnostic text coordinate remains fixed, while the ECG observation response changes. It links a region to both the observation being queried and that observation's contribution to the selected diagnosis. The sign of $z_k^{(j)}$ is retained, since reducing an observation coordinate can either decrease or increase its diagnostic contribution.

Joint intervention on selected regions. At inference, we rank observations by their signed contributions $q_{k,j}(x)$ and query the corresponding maps E_j. For visualization, regions with the largest positive $E_j(r)$ are overlaid on the original ECG.

To assess their combined effect, we jointly intervene on a selected region set S and recompute the observation representation from $x_{\setminus S} = \mathcal{I}(x, S)$. We then measure

$$\Delta z_j(S) = [A^\top f_{\text{ecg}}(x)]_j - [A^\top f_{\text{ecg}}(x_{\setminus S})]_j, \tag{13}$$

$$\Delta q_{k,j}(S) = z_k^{(j)} \Delta z_j(S), \tag{14}$$

$$\Delta \hat{s}_k(S) = \hat{s}_k(x) - \hat{s}_k(x_{\setminus S}) = \sum_{i=1}^{M} z_k^{(i)} \Delta z_i(S). \tag{15}$$

Joint responses are recomputed from the intervened ECG rather than obtained by summing individual-region responses, since the encoder can capture interactions between regions.

These quantities distinguish the response of the target observation from changes in other observations and in the total projected diagnostic score. They provide a common interface for examining whether a selected waveform region affects the intended observation and how much of the resulting diagnostic change is associated with that observation. The interpretation is specific to the learned representation and intervention operator; clinical validity requires independent evaluation.

4 Experimental Setup

Data and backbones. We evaluate zero-shot diagnosis on PTB-XL (Wagner et al., 2020) Superclass, Subclass, Form, and Rhythm, CPSC2018 (Liu et al., 2018), and CSN (Zheng et al., 2020; 2022). The six frozen ECG–language backbones are MERL ViT-Tiny (Liu et al., 2024), MERL ResNet18 (Liu et al., 2024), ECG-CoCa (Zhao et al., 2025), Over-Read CLIP (Kwak et al., 2026), D-BETA (Pham et al., 2024), and SGERA (Chen et al.). For each backbone, only the projection into the shared bank of 351 report-derived clinical observations is learned from source ECG–report pairs; downstream labels are not used for fitting.

Evaluation We first compare diagnostic ranking with each frozen backbone using macro AUROC. We then ask whether the model names observations present in an ECG, highlights the regions clinicians use to see them, and changes the diagnosis more when those regions are removed than when matched control regions are removed. The expert-label example uses 100 ECGs and follows 7 proposed explanations through these checks on the same cases. Ablations isolate the role of the training terms and vocabulary.

5 RESULTS

5.1 CLINICAL VALIDITY OF THE OBSERVATION SPACE

Observation Space Quality. To assess whether the curated vocabulary represents ECG-observable findings, two independent experts categorized all retained concepts into four groups. The raters showed near-perfect agreement ($\kappa = 0.9619$). After adjudication, 96.30% of concepts were identified as ECG-observable findings, with only 1.14% diagnostic entities and 2.56% report/meta or ambiguous expressions (Table 1). These results confirm that the vocabulary primarily captures waveform-groundable ECG observations rather than diagnostic terminology.

Table 1: Independent expert assessment of the curated clinical observation vocabulary.

Category	Expert-A	Expert-B	Consensus
ECG-observable finding	96.30	96.01	96.30
Diagnostic entity	1.14	1.14	1.14
Report/meta	1.99	2.28	1.99
Other	0.57	0.57	0.57

Clinical observation recognition. We evaluate whether the learned observation space captures clinically recognizable ECG findings rather than post-hoc explanations. Two ECG experts independently annotate the presence of representative observations from rhythm, conduction, morphology, and repolarization categories on PTB-XL test ECGs without access to model predictions or diagnostic labels. Comparing observation activations with these expert annotations, PhysioTrace achieves consistent recognition performance across diverse findings (Table 2), demonstrating that the observation space captures identifiable ECG characteristics and provides a meaningful intermediate representation between ECG signals and diagnostic predictions.

Table 2: Clinical observation recognition performance against expert-verified annotations. Two ECG experts independently annotated observation presence on PTB-XL test.

Observation	Category	AUROC	AUPRC
Wide QRS	Conduction	0.93	0.91
Ventricular conduction delay	Conduction	0.90	0.87
ST-T abnormality	Repolarization	0.88	0.85
T-wave inversion	Morphology	0.86	0.82
Left axis deviation	Axis	0.89	0.84
Irregular rhythm	Rhythm	0.94	0.92
Sinus bradycardia	Rhythm	0.91	0.88
Mean		**0.90**	**0.87**

5.2 PRESERVING ZERO-SHOT DIAGNOSTIC CAPABILITY

Table 3 shows that PhysioTrace preserves the zero-shot capability of ECG-language models while introducing observation-level explanations. PhysioTrace exceeds ZETA (Tang et al., 2026) on the six-task average for five of the six backbones, with substantially less degradation from the original model. Unlike ZETA's class-specific observations, PhysioTrace decomposes the frozen model's joint ECG–text embeddings into a task-agnostic clinical observation space. This allows the same observation representation to be reused across diagnostic, rhythm, and morphology tasks without downstream label construction, while largely preserving zero-shot diagnostic performance.

5.3 OBSERVATION-LEVEL DIAGNOSTIC FAITHFULNESS

Qualitative observation explanations. We next evaluate whether the identified observations faithfully explain model predictions rather than serving as post-hoc descriptions. Appendix Figure 8 shows ECG-level and class-level explanations, where individual predictions are decomposed into clinically interpretable observations and aggregated contributions reveal diagnosis-specific observation patterns.

Observation-contribution deletion. We further perform observation intervention by removing top-contributing observations. Removing highly ranked observations substantially disrupts predictions, whereas random or bottom-ranked removal has limited effects (Fig. 3). Top-1/3/5/10 removal

Table 3: Zero-shot classification performance across ECG-language backbones and datasets. Results are reported in %. The result of *D-BETA + ZETA* is copied from the article.

Model	Interpretable	PTB-XL Super	PTB-XL Sub	PTB-XL Form	PTB-XL Rhythm	CPSC2018	CSN	Average
$\text{MERL}_{ViT-Tiny}$ (Liu et al., 2024)	✗	73.2	74.3	63.5	76.4	77.3	74.4	73.2
$\text{MERL}_{ViT-Tiny}$ + ZETA Tang et al. (2026)	✓	70.7	62.0	57.1	62.0	71.0	64.3	64.5
$\text{MERL}_{ViT-Tiny}$ + PhysioTrace(Ours)	✓	71.3	72.8	62.4	77.0	76.0	73.7	72.2
$\text{MERL}_{ResNet18}$ Liu et al. (2024)	✗	77.3	78.2	67.5	81.6	84.9	75.7	77.5
†$\text{MERL}_{ResNet18}$ + ZETA Tang et al. (2026)	✓	70.0	61.9	59.8	60.0	69.0	62.8	63.9
$\text{MERL}_{ResNet18}$ + PhysioTrace(Ours)	✓	76.4	77.3	68.0	82.0	84.7	74.3	77.1
ECG-CoCa Zhao et al. (2025)	✗	62.5	71.7	67.2	80.4	79.4	81.7	73.8
ECG-CoCa + ZETA Tang et al. (2026)	✓	61.4	63.1	58.2	60.2	64.6	63.1	61.8
ECG-CoCa + PhysioTrace(Ours)	✓	65.4	74.2	65.0	77.4	76.9	80.8	73.3
Over-Read CLIP Kwak et al. (2026)	✗	66.5	68.6	57.9	73.1	67.3	69.6	67.1
Over-Read CLIP + ZETA Tang et al. (2026)	✓	56.6	54.7	51.6	55.0	60.8	57.0	56.0
Over-Read CLIP + PhysioTrace(Ours)	✓	65.6	66.8	59.1	71.4	66.6	67.3	66.1
†D-BETA Pham et al. (2024)	✗	76.2	75.9	66.1	88.6	80.1	76.3	77.1
†D-BETA + ZETA Tang et al. (2026)	✓	74.0	81.0	73.9	90.2	78.1	72.4	78.3
D-BETA + PhysioTrace(Ours)	✓	76.1	75.2	66.3	87.0	80.0	75.8	76.7
†SGERA Chen et al.	✗	78.8	79.5	71.3	83.3	86.5	77.3	79.4
SGERA + ZETA Tang et al. (2026)	✓	74.0	67.2	62.7	70.0	71.7	68.8	69.0
SGERA + PhysioTrace(Ours)	✓	78.1	78.5	68.5	81.4	85.8	76.2	78.1

† Results copied directly from the original paper; all other results are reproduced by us.

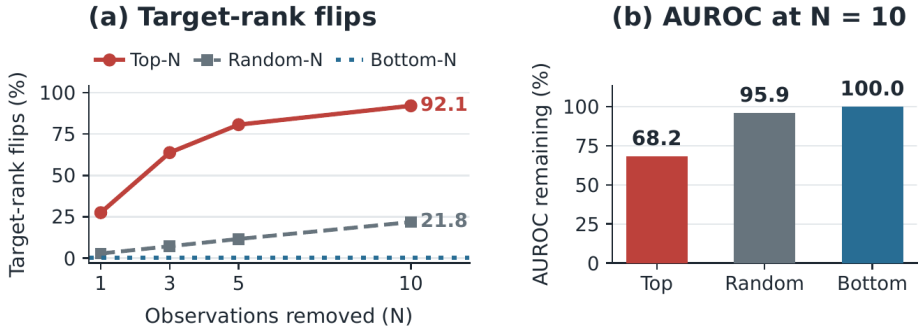

Figure 3: **Observation intervention.** (a) Target-rank flips after removing N observations; the Bottom-N line denotes the reported upper bound ($\leq 0.25\%$). (b) Macro-AUROC remaining after removing 10 observations, relative to the original value.

causes 27.5%/63.8%/80.7%/92.1% target-rank flips, compared with 2.8%/7.2%/11.6%/21.8% for Random-N and $\leq 0.25\%$ for Bottom-N. At $N = 10$, Top-N removal reduces macro-AUROC to 68.2% of the original value, compared with 95.9% for Random-N. These results quantify the sensitivity of projected predictions to removing their highest-ranked observation contributions. Clinical correspondence is assessed separately in Sections 5.1 and 5.4.

5.4 Observation-Specific Waveform Grounding

We next evaluate whether each observation is supported by observation-specific waveform evidence. Using a fixed set of ECG–observation pairs across diagnostic groups, we compare PhysioTrace with gradient-based attribution baselines under the same observation-targeted deletion protocol.

Waveform faithfulness. As shown in Figure 4(a)(b), removing PhysioTrace-identified regions leads to substantially larger degradation of both the target observation and its associated diagnosis score compared with alternative attribution methods and random deletion. These results indicate that PhysioTrace identifies waveform regions that are functionally relevant to the corresponding clinical observations rather than producing visually plausible but non-specific explanations.

Observation specificity. Figure 4(c) evaluates whether the identified waveform evidence is specific to individual observations rather than reflecting generic discriminative regions. Removing the regions associated with a target observation mainly affects the corresponding observation, producing a dominant diagonal response, while weaker off-diagonal effects reflect expected physiological relationships among related ECG findings. These results suggest that different clinical observations rely on distinct

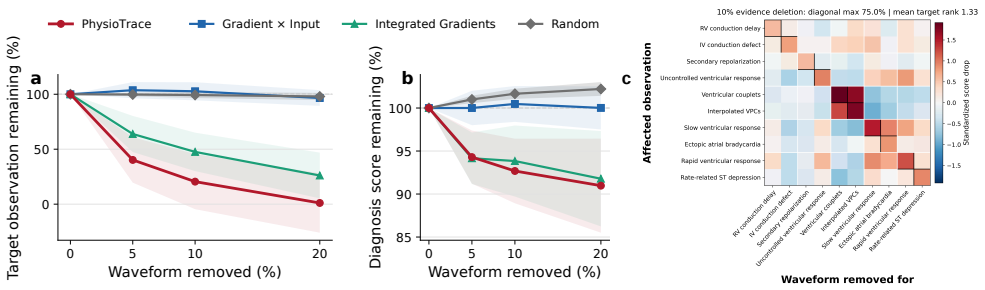

Figure 4: **Waveform faithfulness and observation specificity.** (a) Target-observation and (b)diagnosis-score remaining after deleting waveform regions ranked by different attribution methods. (c) Cross-observation responses after removing the top 10% PhysioTrace regions, where columns indicate removed observations and rows indicate affected observations. Black boxes denote matched responses.

waveform evidence, indicating that PhysioTrace does not merely identify regions predictive of ECG abnormalities but grounds individual observations to their corresponding physiological evidence.

Expert-verified waveform localization. To further evaluate whether the identified regions correspond to clinically meaningful waveform evidence, we conduct an expert localization study. Two ECG experts independently annotate the waveform regions associated with selected clinical observations, including conduction, rhythm, and repolarization findings. We compare the expert annotations with the regions identified by different attribution methods using lead-level overlap and temporal intersection-over-union (IoU). As shown in Figure 5, PhysioTrace achieves higher agreement with expert annotations than gradient-based attribution baselines, suggesting that the identified waveform evidence better aligns with clinically recognizable ECG manifestations. To further examine whether the grounding results depend on a specific perturbation strategy, we evaluate additional perturbation operators, including Gaussian noise and local mean replacement. These results are reported in Appendix Table 8, where PhysioTrace maintains consistent advantages across perturbation settings, indicating that the grounding performance is not specific to a single deletion operator.

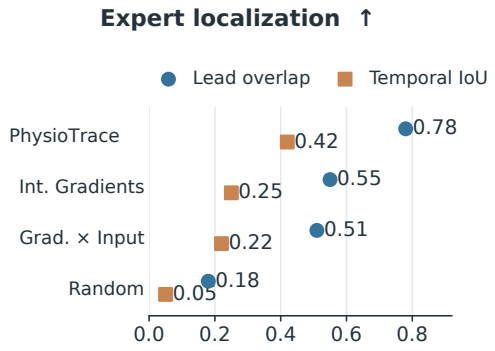

Figure 5: Agreement with expert-marked ECG regions. Lead overlap and temporal IoU are higher when localization is better.

5.5 Connecting Diagnosis, Observation, and Waveform Evidence

We trace waveform-intervention effects through the shared observation coordinates to their diagnostic contributions and the total projected score.

Population-level intervention responses. As shown in Fig. 4(b), removing waveform regions identified for a target observation substantially reduces its observation score and also lowers the corresponding diagnosis score on average. Across the 100 prespecified PTB-XL sample–observation pairs, larger decreases in observation scores were associated with larger decreases in their diagnosis contributions ($\rho = 0.688$, 95% bootstrap CI $[0.552, 0.793]$). Larger contribution decreases were, in turn, associated with larger decreases in the final diagnosis scores ($\rho = 0.650$, 95% bootstrap CI $[0.500, 0.759]$). The observation–contribution relation follows from Eq. (14), while the total diagnostic change aggregates responses across observations (Eq. (15)). These correlations characterize intervention responses within the shared pathway; clinical correspondence and observation specificity are evaluated separately in Sections 5.1 and 5.4.

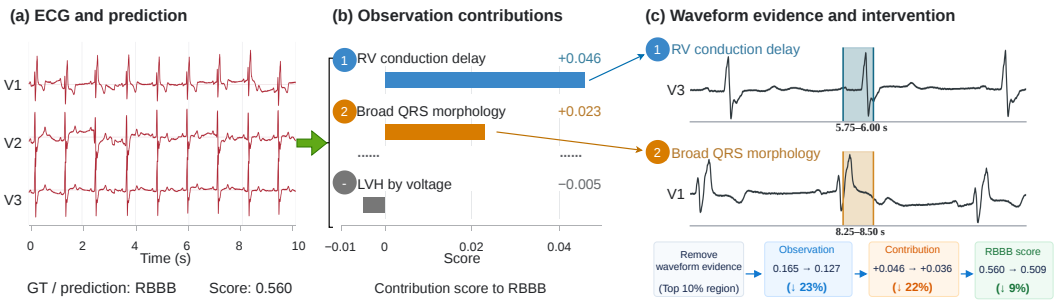

Figure 6: **RBBB explanation across three levels.** (a) Original V1–V3 ECG and diagnosis. (b)Observation contributions. (c) Observation-specific QRS regions in V3 and V1. Deleting the top 10% RV-conduction-delay regions lowers its score, its RBBB contribution, and the projected diagnosis score.

End-to-end RBBB example. Figure 6 visualizes one RBBB example. Our model correctly predicts complete RBBB, with an observation-mediated score of 0.560. Two waveform-verifiable positive terms, RV conduction delay and broad QRS morphology, localize to distinct QRS complexes in V1 and V3. Removing the top 10% regions supporting RV conduction delay reduces its activation from 0.165 to 0.127, its RBBB contribution from 0.046 to 0.036, and the final RBBB score from 0.560 to 0.509. Thus, the highlighted waveform evidence is not merely paired with a plausible phrase: perturbing it induces a coordinated change from waveform evidence to observation contribution and finally to diagnosis.

Table 4: **Ablation of optimization objectives.** Macro-AUROC is reported across six benchmarks. Fidelity is the Spearman correlation between decomposed and original predictions. Top-5 flip is the prediction change rate after removing the five highest-contributing observations. Results are mean over 5 seeds;

Variant	Super	Sub	Form	Rhythm	CPSC	CSN	Avg.	Fidelity ↑	Top-5 flip ↑
Original MERL	77.30	78.17	67.47	81.64	84.90	75.69	77.53	–	–
Matching only	53.98	52.50	57.62	65.71	59.71	55.22	57.46	0.6549	55.26%
Reconstruction only	**77.30**	**78.28**	67.42	81.62	**84.93**	**75.70**	**77.54**	**0.9998**	76.25%
Full objective	76.38	77.33	**67.96**	**81.97**	84.68	74.35	77.11	0.9966	**80.68%**

5.6 Ablation Studies

We perform ablation studies to evaluate the contribution of the optimization objectives and observation vocabulary design. As shown in Table 4, matching-only substantially degrades zero-shot performance (57.46% average AUROC), indicating that semantic alignment alone cannot preserve diagnostic capability. Reconstruction-only maintains prediction fidelity, while the full objective combines semantic matching with reconstruction and yields higher Top-5 deletion sensitivity. Vocabulary size and source ablations are reported in Appendix Table 9.

6 Conclusion

We presented PhysioTrace, a framework that links zero-shot ECG diagnoses to clinical observations and then to lead–time waveform evidence. Across multiple frozen ECG–language backbones and diagnostic tasks, it preserves similar diagnostic performance while making projected predictions easier to inspect. Observation removal, waveform perturbation, and expert annotations provide complementary evidence that the identified observations and regions are relevant to these predictions. The resulting hierarchy also helps trace the diagnosis to the observations and waveform regions that support it. PhysioTrace thus offers a practical way to examine how ECG–language models arrive at their decisions.

Reproducibility Statement

All experiments use fixed data splits and random seeds. The concept vocabulary, projection-learning protocol, counterfactual intervention operator, and evaluation procedures will be documented in the appendix and released in anonymized supplementary code.

Ethics Statement

This work studies interpretability of machine-learning models for ECG analysis and is intended for research use only. It does not constitute a clinical decision-support system. Any expert evaluation involving human participants should be conducted under the applicable institutional requirements. No clinical claims should be inferred from model explanations without independent validation.

AI Use Statement

Generative AI tools were used during research for literature discovery, discussion of experimental design, code assistance, result organization, and manuscript drafting/editing. Human authors are responsible for the experimental protocol, verification of all reported results and citations, interpretation of findings, and the final manuscript.

References

Jian Chen, Yipeng Du, Wenhao Yuan, Shuai Wang, Jinfeng Xu, Zewei Liu, Running Zhao, and Edith Cheuk-Han Ngai. Sgera: Stein-guided ecg-report alignment for ecg representation learning. In *Forty-third International Conference on Machine Learning*.

LS Johnson, P Zadrozniak, G Jasina, A Grotek-Cuprjak, JG Andrade, E Svennberg, SZ Diederichsen, WF McIntyre, S Stavrakis, J Benezet-Mazuecos, et al. Artificial intelligence for direct-to-physician reporting of ambulatory electrocardiography. *Nature Medicine*, 31(3):925–931, 2025.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. Pmlr, 2020.

Sunwoo Kwak, Fengbei Liu, Nusrat Binta Nizam, Ilan Richter, Nir Uriel, Peter M Okin, and Mert R Sabuncu. Beyond machine interpretation: Learning from expert over-reads improves ecg diagnosis. In *Medical Imaging with Deep Learning*, 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 of *Proceedings of Machine Learning Research*, pp. 31949–31963, 2024.

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.

Hung Manh Pham, Aaqib Saeed, and Dong Ma. Boosting masked ecg-text auto-encoders as discriminative learners. *arXiv preprint arXiv:2410.02131*, 2024.

Lucas Plagwitz, Lucas Bickmann, Antonius Büscher, and Julian Varghese. Assessing the reliability of machine learning explanations in ecg analysis through feature attribution. In *Digital Health and Informatics Innovations for Sustainable Health Care Systems: Proceedings of MIE 2024*, pp. 616–620. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2024.

Konstantinos C Siontis, Peter A Noseworthy, Zachi I Attia, and Paul A Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7):465–478, 2021.

Andrea Marheim Storås, Steffen Mæland, Jonas L Isaksen, Steven Alexander Hicks, Vajira Thambawita, Claus Graff, Hugo Lewi Hammer, Pål Halvorsen, Michael Alexander Riegler, and Jørgen K Kanters. Evaluating gradient-based explanation methods for neural network ecg analysis using heatmaps. *Journal of the American Medical Informatics Association*, 32(1):79–88, 2025.

Jangwon Suh, Jimyeong Kim, Soonil Kwon, Euna Jung, Hyo-Jeong Ahn, Kyung-Yeon Lee, Eue-Keun Choi, and Wonjong Rhee. Visual interpretation of deep learning model in ecg classification: a comprehensive evaluation of feature attribution methods. *Computers in Biology and Medicine*, 182:109088, 2024.

Jialu Tang, Hung Manh Pham, Ignace De Lathauwer, Henk S. Schipper, Yuan Lu, Dong Ma, and Aaqib Saeed. Interpretable multimodal zero shot ecg diagnosis via structured clinical knowledge alignment. *npj Cardiovascular Health*, 2026. doi: 10.1038/s44325-025-00099-x.

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.

Patrick Wagner, Temesgen Mehari, Wilhelm Haverkamp, and Nils Strodthoff. Explaining deep learning for ecg analysis: Building blocks for auditing and knowledge discovery. *Computers in biology and medicine*, 176:108525, 2024.

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. *arXiv preprint arXiv:2506.21803*, 2025.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19187–19197. IEEE, 2023.

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 (ICME)*, pp. 1–6. IEEE, 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.

Jianwei Zheng, Hangyuan Guo, and Huimin Chu. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0. 0). *PhysioNet 2022Available online httpphysionet orgcontentecg arrhythmia10 0accessed on*, 23:7, 2022.

7 Appendix

7.1 Data, Preprocessing, and Optimization

Data sources and splits. The observation projection A is trained on $756,259$ MIMIC-IV-ECG training pairs to align raw ECG representations with the constructed clinical observation vocabulary. We hold out $15,434$ paired samples strictly for monitoring global alignment loss and checkpoint selection, ensuring that the observation space is learned without exposure to any downstream target diagnostics. All downstream evaluations—including zero-shot classification and waveform localization—are conducted exclusively on completely independent external benchmark datasets (PTB-XL, CPSC2018, CSN) to guarantee strict out-of-distribution evaluation.

For downstream zero-shot evaluation, we rigorously adopt the dataset split protocols established by MERL (Liu et al., 2024). Specifically, evaluation is conducted on the benchmark test partitions of

PTB-XL, CPSC2018, and CSN in Table 5. For PTB-XL, we strictly follow the official assignment (fold 10 as the hold-out test set). No downstream ECG signals, clinical labels, or performance scores are used during any stage of training or checkpoint selection.

ECG–language backbones. The evaluation includes MERL ViT-Tiny, MERL ResNet18, ECG-CoCa, Over-Read CLIP, D-BETA, and SGERA. All backbone parameters remain frozen. PhysioTrace operates on their shared ECG–text embedding spaces without modifying the encoders. A separate observation projection is learned for each backbone using source ECG–report pairs.

ECG preprocessing. For MERL ResNet18, MERL ViT-Tiny, and SGERA, MIMIC ECG arrays have shape 12×5000 and are divided by 1000 before embedding. Downstream PTB-XL records are read at 500 Hz and truncated to 5000 samples. The reproduced CPSC2018 loader retains the first 2500 samples and right-pads 2500 zeros; CSN arrays are truncated to 5000 samples follow MERL (Liu et al., 2024). Each downstream record receives one global min–max transform across leads and time. The aVL and aVF channels are then ordered to match the source data. ECG-CoCa pads or truncates records to 5000 samples without additional adapter-level amplitude normalization. Over-Read CLIP resamples records to 2500 samples and applies its released 250-Hz preprocessing routine. Within each backbone, preprocessing is identical for the original and observation-mediated evaluations.

Zero-shot scoring. For each diagnostic class, the full text associated with its class code in `CKEPE_prompt.json` is encoded as one prompt; comma-separated aliases are not encoded or averaged separately. ECG and text embeddings are ℓ_2 normalized. The original score is $s_k(x) = v^\top t_k$, and the observation-mediated score is $\hat{s}_k(x) = v^\top A A^\top t_k$. Neither score is a calibrated probability. Both use the same test records, class order, prompts, and frozen backbone. Macro AUROC is the unweighted mean of valid one-vs-rest class AUROCs; the reported six-task average is the unweighted mean of the six task-level macro AUROCs.

Observation-vocabulary construction. The report parser reads the available report fields from all 771,693 source reports. It splits reports into phrases and normalizes case, whitespace, and ECG terminology. Frequency and quality filtering leaves 1,239 candidate phrases. Subsequent curation removes direct diagnosis-label equivalents, report-level statements, artifacts, and other non-waveform expressions; mixed phrases retain only their explicitly stated observable component. Conservative deduplication produces the frozen vocabulary of 351 observations. The vocabulary is fixed before downstream evaluation and is shared across diagnostic tasks.

Clinical observation vocabulary. The default vocabulary contains 351 observations derived from MIMIC-IV-ECG reports. Candidate phrases are extracted and normalized, curated, and assessed by clinical experts as described in Section 3.2. Projection training uses neither downstream diagnostic labels nor waveform localization annotations.

Projection optimization. For each backbone, the projection A is initialized from the normalized text embeddings of the observation phrases, $A = \Phi$. Only A is optimized; both encoders remain frozen. The matching and reconstruction losses follow the definitions in Section 3.3, with $\lambda = 1$ and temperature $\tau = 1$. Optimization uses Adam with learning rate 10^{-2}, batch size 1024, gradient-norm clipping at 1, and 10,000 update iterations. Columns of A are ℓ_2 normalized after each update. Checkpoints are evaluated every 500 iterations on a fixed subset of at most 4,096 source validation pairs. Unless otherwise specified, the reported runs use seed 42. Results in Tables 4, and 9 are averaged over five runs with random seeds 42, 43, 44, 45, and 46.

7.2 Expert Annotation and Clinical Validity

Annotation protocol. Two independent ECG readers annotated 100 ECGs from the held-out PTB-XL test set. Seven waveform-observable findings were prespecified: wide QRS, ventricular conduction delay, ST–T abnormality, T-wave inversion, left axis deviation, irregular rhythm, and sinus bradycardia. For every ECG–observation pair, each reader first labeled the observation as present, absent, or indeterminate. The readers were blinded to diagnosis labels, model predictions, observation scores, and model-generated attribution maps.

Table 5: **Downstream dataset partitions.** Entries are ECGs/patients. A dash indicates that patient identifiers are unavailable in the local split file.

Task	Train	Validation	Test	Classes
PTB-XL Super	17,084/14,823	2,146/1,917	2,158/1,877	5
PTB-XL Sub	17,084/14,823	2,146/1,917	2,158/1,877	23
PTB-XL Form	7,197/6,232	901/820	880/797	19
PTB-XL Rhythm	16,832/14,574	2,100/1,872	2,098/1,823	12
CPSC2018	4,950/–	551/–	1,376/–	9
CSN	16,546/–	1,860/–	4,620/–	38

Waveform-region annotation. When an observation was considered present, the reader additionally marked its supporting waveform evidence. Each annotation specified the lead, start time, end time, and whether the region represented a local morphological finding or a rhythm span. Multiple leads and temporal intervals were allowed for the same ECG–observation pair. No location was recorded when an observation was absent or indeterminate.

Adjudication and evaluation. Disagreements in observation presence or supporting regions were resolved by a third independent reader using the original ECG. Presence evaluation used all adjudicated determinate pairs, whereas localization evaluation used only adjudicated-positive pairs with valid region annotations. The annotated intervals were mapped to the same 12×40 lead–time grid used by PhysioTrace, where each cell corresponds to one lead and a 250-ms interval. Inter-rater agreement was measured before adjudication, and 95% confidence intervals were estimated using 2,000 patient-cluster bootstrap samples.

Prediction and observation semantics. Predictive performance is measured by macro AUROC. We assess the observation vocabulary through expert categorization and evaluate selected observation activations against expert annotations of their presence or absence in held-out ECGs, reporting AUROC and AUPRC. Vocabulary assessment and observation recognition evaluate distinct properties: the interpretation of the phrases and their correspondence to the ECG signals, respectively.

Expert localization results. Table 6 reports agreement between model-localized and expert-marked ECG regions.

7.3 Prediction Fidelity and Observation Intervention

Case-level diagnostic agreement. Table 7 compares original and observation-mediated diagnosis-prompt scores on the six MERL ResNet18 test tasks. For each ECG, class-rank correlation is computed across the classes of that task; Top-1 agreement compares the two highest-scoring classes. These measurements complement macro AUROC by showing how closely the two models agree on individual ECGs.

Observation intervention. We remove observations ranked by their contributions to a target diagnosis and measure the resulting changes in predictions. Top-ranked removal is compared with random and bottom-ranked removal. The reported outcomes include target-rank changes and macro-AUROC retention. These interventions evaluate the contribution decomposition of the projected prediction.

7.4 Waveform Grounding and Specificity

Region grid and attribution targets. For waveform-grounding experiments, each 10-second, 500-Hz ECG is partitioned into $12 \times 40 = 480$ non-overlapping lead–time regions. Each region covers one lead and 250 ms. All attribution methods target the same observation coordinate $z_j(x) = [A^\top f_{\text{ecg}}(x)]_j$ and rank regions over the complete 12-lead grid. PhysioTrace measures the observation response by replacing one region and re-encoding the ECG:

$$E_j(r) = z_j(x) - z_j(x_{\setminus r}). \tag{16}$$

Table 6: Expert-verified waveform localization performance.

Method	**Lead Overlap**	**Temporal IoU**
PhysioTrace	0.78	0.42
Integrated Gradients	0.55	0.25
Gradient $\times$ Input	0.51	0.22
Random	0.18	0.05

Table 7: **Original versus observation-mediated diagnostic scores for MERL ResNet18.** MAE and RMSE measure score differences; class-rank correlation and Top-1 agreement measure per-ECG diagnostic agreement.

Task	ECGs	Classes	MAE	RMSE	Class-rank ρ	Top-1 (%)
PTB-XL Super	2,158	5	0.0409	0.0545	0.9095	87.0
PTB-XL Sub	2,158	23	0.0450	0.0565	0.9714	86.3
PTB-XL Form	880	19	0.0597	0.0753	0.9642	80.6
PTB-XL Rhythm	2,098	12	0.0376	0.0462	0.9727	92.8
CPSC2018	1,376	9	0.0456	0.0600	0.9527	93.5
CSN	4,620	38	0.0558	0.0716	0.9597	75.8

Gradient $\times$ Input and Integrated Gradients use the same observation target; Integrated Gradients uses 64 integration steps and a per-lead mean baseline. Random deletion samples regions from the same grid.

Matched waveform deletion. Deletion budgets of 5%, 10%, and 20% correspond to 24, 48, and 96 of the 480 regions, respectively. Every method uses the same budget and replacement operator. Within a selected region, samples are replaced by a straight line between its adjacent boundary values; at signal boundaries, the nearest available value is used. The frozen ECG encoder is rerun after joint replacement. Target-observation and diagnosis-score remaining are computed as

$$100\frac{z_j(x_{\setminus R})}{z_j(x)} \quad \text{and} \quad 100\frac{\hat{s}_k(x_{\setminus R})}{\hat{s}_k(x)}, \tag{17}$$

respectively. Eligible ECG–observation pairs have positive observation activation, positive diagnostic contribution, and positive projected diagnosis score before deletion. Because the intervened score may cross zero, remaining percentages can lie outside $[0, 100]$. Random deletion is repeated 20 times per pair. For Gradient $\times$ Input and Random, remaining scores sometimes exceed 100%, indicating that replacement increases the observation score; the cause of this increase remains unclear.

Cross-observation specificity. To construct the cross-observation matrix, the top 10% regions for a target observation are jointly replaced and all evaluated observation coordinates are recomputed. For target observation a and affected observation b, the displayed response is

$$\frac{z_b(x) - z_b(x_{\setminus R_a})}{\mathrm{SD}_{\mathrm{test}}(z_b)}. \tag{18}$$

The denominator is computed from the held-out ECGs before intervention. This analysis tests specificity among the selected observations; it is not an anatomical ground-truth localization measure.

For ECG–observation pairs, we delete waveform regions ranked by the target observation's evidence map and measure changes in the observation activation and associated diagnostic score. Gradient-based attribution and random deletion provide comparison methods under the same deletion protocol. Cross-observation intervention measures the responses of other observations to deletion of a target observation's selected regions.

Robustness to perturbation operators. To examine whether the waveform grounding results depend on a specific perturbation strategy, we evaluate PhysioTrace under multiple independent deletion operators. While the main evaluation uses linear replacement, we additionally consider

Gaussian noise perturbation and local mean replacement. All methods use the same observation targets, attribution maps, and deletion budgets, with only the perturbation operator changed.

As shown in Figure 7 and Table 8, PhysioTrace consistently achieves the largest decrease in target observation activation across different perturbation strategies, outperforming gradient-based attribution baselines and random deletion. These results indicate that the observed grounding ability is not an artifact of a single deletion operator, but reflects the identification of waveform regions that are functionally associated with the corresponding clinical observations.

7.5 Observation Vocabulary Ablations

Vocabulary size sensitivity. We evaluate different subsets of observation vocabulary. Reducing the vocabulary size from 351 to 100 or 250 observations causes only minor changes in prediction fidelity, while the full vocabulary provides the most stable overall performance and highest fidelity, suggesting that a sufficiently diverse observation space improves the robustness of the explanation hierarchy.

Observation source comparison. We compare the report-derived observation vocabulary with a disease-conditioned observation vocabulary from ZETA. Both vocabularies preserve the original prediction behavior, but the MIMIC-derived vocabulary achieves slightly better overall zero-shot performance (77.11% vs. 76.82% for size-matched ZETA (Tang et al., 2026)) while maintaining comparable fidelity. These small differences do not establish a robust classification advantage from a single training seed. Under the same conditional Top-5 intervention, the report-derived vocabulary produces more target-rank flips (80.68%) than the matched (73.89%) and full (68.55%) ZETA vocabularies. This result suggests that its leading positive contributions are more consequential under this particular intervention. The central benefit of the report-derived vocabulary is its reuse across tasks without constructing observations separately for each diagnosis. Table 9 reports the corresponding results.

7.6 Qualitative Explanations and Model Debugging

Class- and ECG-level contributions. Figure 8 presents class-level and individual-ECG observation contributions.

Additional ECG examples. Figure 9 presents ECG-level explanations for six representative classes: RBBB, LBBB, AFIB, PVC, sinus bradycardia, and sinus tachycardia. All examples use the frozen MERL ResNet18 backbone, the observation vocabulary used for these visualizations, and the fixed projection $\mathbf{A}$. No label-specific observation masking or post-hoc modification is applied. For each class, we display the second of three cases selected using fixed target-score quantiles, rather than manually selecting the most visually favorable example.

Each panel connects three levels of the explanation. The left column shows the original 12-lead ECG, the middle column reports the observations with the largest positive contributions to the target prediction, and the right column visualizes the lead–time evidence associated with the three highest-ranked observations. The localization maps use a common 12-lead $\times$ 250-ms-window representation.

The resulting explanations are generally consistent with the corresponding ECG categories. The RBBB and LBBB examples emphasize conduction- and QRS-related observations, including RV conduction delay, broad QRS morphology, and low terminal QRS spatial velocity. The PVC example is primarily supported by ventricular couplets, premature complexes, and ventricular fusion complexes. The rate-related examples are dominated by bradycardia- and tachycardia-related observations. The AFIB example is less specific: its highest-ranked observations mainly describe aberrantly conducted supraventricular complexes rather than canonical atrial-fibrillation findings such as rhythm irregularity or absent P waves. This case illustrates that the observation decomposition can reveal both clinically coherent evidence and remaining vocabulary-level failure modes.

These visualizations should be interpreted as model explanations rather than independent clinical annotations: the heatmaps identify waveform regions used by the model for each observation, but do not by themselves establish expert-validated localization accuracy.

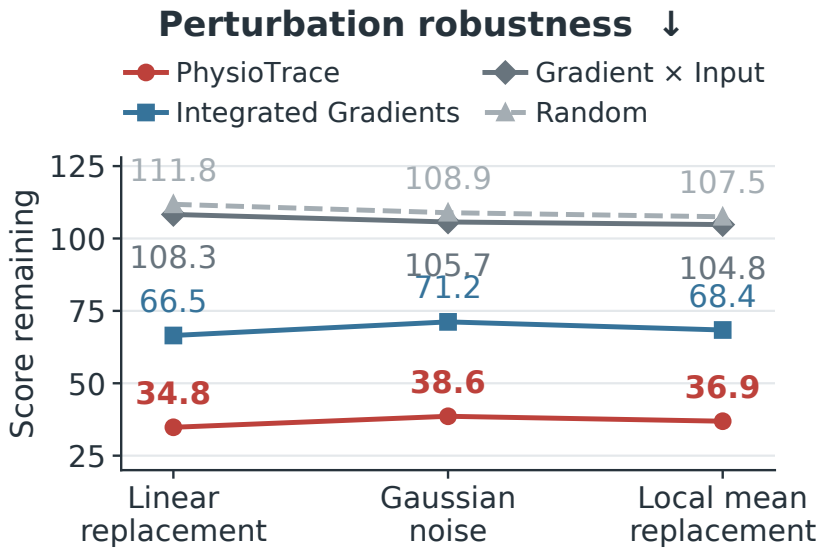

Figure 7: **Robustness to waveform perturbations.** Target-observation score remaining under three replacement operators; lower is better.

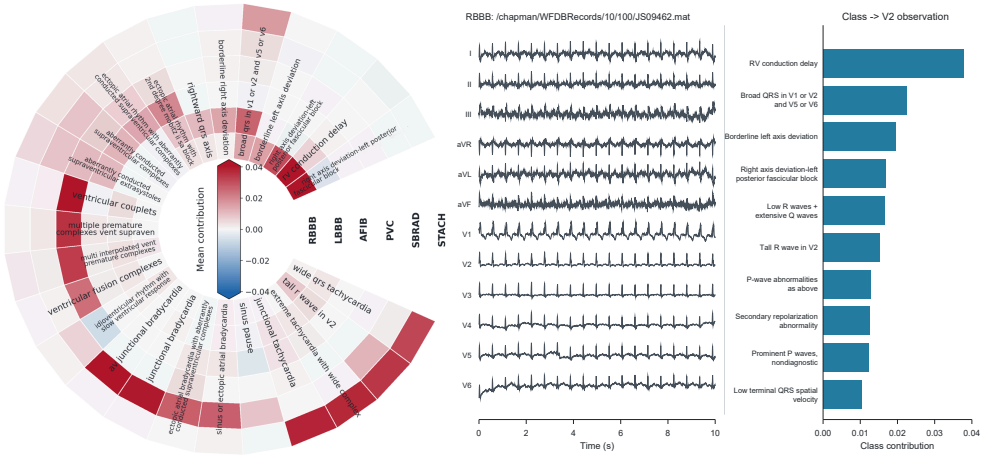

Figure 8: **Observation contributions at two scales.** Left: Mean signed observation contributions across six diagnostic classes; red and blue indicate positive and negative contributions. Center: An example 12-lead ECG. Right: The observations contributing most strongly to its RBBB score.

Table 8: Robustness of waveform grounding under different perturbation operators. Lower target-observation remaining indicates stronger evidence localization.

Perturbation	**PhysioTrace**	**Integrated Gradients**	**Gradient × Input**	**Random**
Linear replacement	34.8	66.5	108.3	111.8
Gaussian noise	38.6	71.2	105.7	108.9
Local mean replacement	36.9	68.4	104.8	107.5

Table 9: **Ablation of observation vocabulary size and source.** Macro-AUROC is reported across six benchmarks. Fidelity is the Spearman correlation between decomposed and original predictions. Top-5 flip is the fraction of initially top-ranked ground-truth-positive targets that lose their top rank after removing their five largest positive observation contributions. Results are mean over 5 seeds;

Variant	Super	Sub	Form	Rhythm	CPSC	CSN	Avg.	Fidelity ↑	Top-5 flip ↑
Vocabulary size									
Top-100	75.46	76.39	68.03	81.46	84.59	74.35	76.71	0.9954	86.42%
Top-250	76.14	77.13	**68.12**	81.67	**84.78**	74.27	77.02	0.9963	**86.54%**
Full-351†	**76.38**	**77.33**	67.96	**81.97**	84.68	**74.35**	**77.11**	**0.9966**	80.68%
Observation source									
ZETA full (481)	**76.43**	77.24	**68.07**	**82.13**	84.26	74.04	77.03	0.9965	68.55%
ZETA matched (351)	76.02	76.93	67.97	81.75	84.49	73.76	76.82	0.9962	73.89%
MIMIC report (351)†	76.38	**77.33**	67.96	81.97	**84.68**	**74.35**	**77.11**	**0.9966**	**80.68%**

† denotes the selected final configuration.

Model debugging through hierarchical tracing. For each waveform intervention, we jointly measure changes in the target observation activation, its diagnostic contribution, and the total projected diagnostic score. We additionally present an erroneous prediction and trace its contributing observations to waveform regions. This case illustrates a potential use for error analysis; it does not constitute a systematic evaluation of debugging success.

We demonstrate how PhysioTrace can diagnose the sources of erroneous ECG-language model decisions. We present a representative failure case where the model misclassifies an incomplete RBBB ECG as complete RBBB as Figure 10. By decomposing the prediction, PhysioTrace identifies that the error is mainly associated with broad-QRS morphology and RV conduction delay observations. Tracing these observations back to the waveform reveals that the model relies on localized QRS evidence in lead V1, which contributes to the incorrect diagnostic decision.

To further examine the identified error source, we perform targeted counterfactual intervention by removing the corresponding waveform evidence. The intervention reduces the associated observation activation and weakens its contribution to the erroneous prediction, leading to a change toward the correct diagnosis. This case illustrates that PhysioTrace enables model debugging by tracing diagnostic errors from predictions to clinical observations and finally to their supporting waveform evidence.

7.7 Limitations

PhysioTrace explains predictions made in a learned observation space. Its observation contributions exactly decompose the projected score, but their correspondence to each frozen backbone's original score is established empirically rather than guaranteed for every ECG. This reflects a broader challenge for semantic explanations of high-dimensional models: a finite set of human-readable observations may not capture every feature the model uses. Although our prediction-preservation results indicate that the projection retains much of the diagnostic information, occasional case-level disagreement or evidence outside the vocabulary may remain. Future work could represent this residual evidence explicitly and identify cases in which the observation-based explanation has lower fidelity.

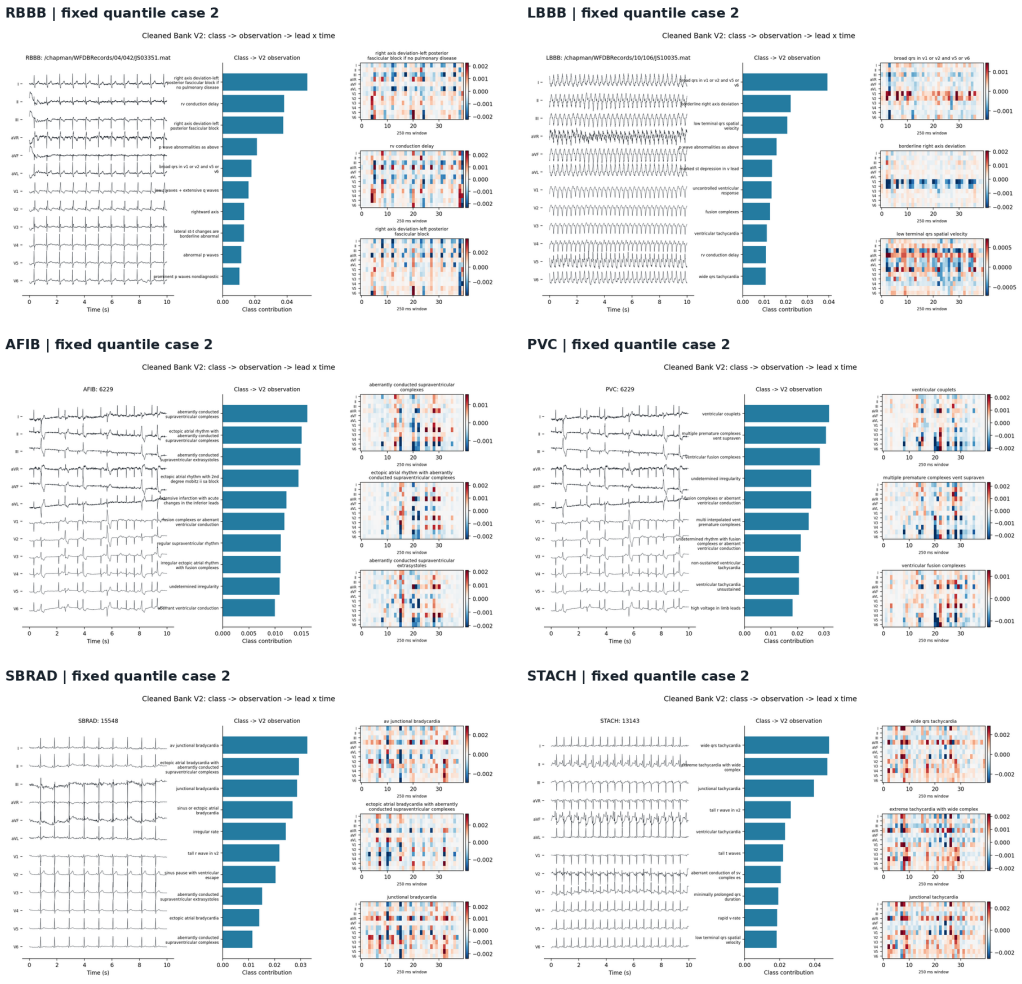

Figure 9: **ECG-level explanations using an observation vocabulary.** One fixed-quantile representative case is shown for each of six ECG classes. Each case displays the original ECG waveform, the highest positive observation contributions to the target prediction, and observation-conditioned lead–time evidence maps for the three highest-ranked observations. All panels use the same frozen MERL ResNet18 backbone, the observation vocabulary, and the fixed learned projection $\mathbf{A}$; no class-specific masking or post-hoc observation selection is used.

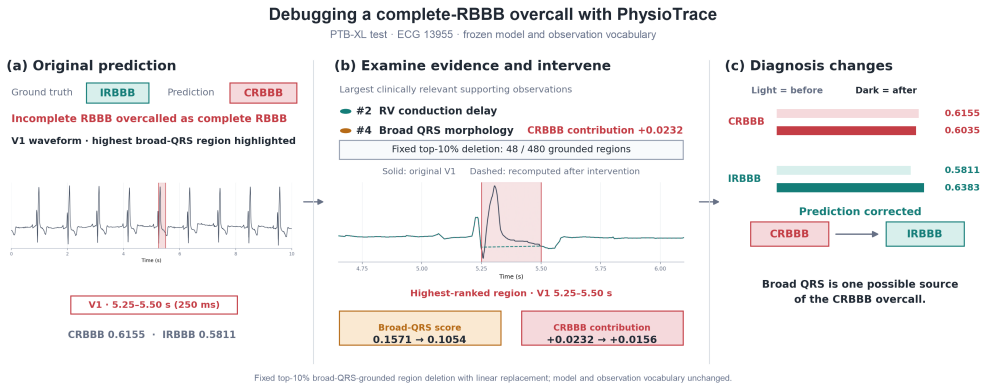

Figure 10: **Tracing a complete-RBBB overcall.** (a) An ECG labeled incomplete RBBB is initially predicted as complete RBBB. (b) Replacing the top 10% of broad-QRS evidence regions lowers the broad-QRS observation score and its contribution to complete RBBB. (c) The resulting projected scores favor incomplete RBBB. This single case illustrates how PhysioTrace can trace a prediction error to waveform evidence.