ONE REPRESENTATION FOR DETECTION AND INTERPRETATION: LANGUAGE-ANCHORED UNIFIED ECG REPRESENTATION LEARNING VIA SIGNAL, IMAGE, AND CLINICAL TEXT

Anonymous authors
Paper under double-blind review

ABSTRACT

Electrocardiogram (ECG) representation is a core for computer-aided diagnosis of cardiovascular diseases (CVDs). The existing ECG-Text alignment-based methods are a promising paradigm for ECG representation learning. However, the learned representations struggle to concurrently execute both disease detection and interpretation in the diagnosis of CVDs, as proficiently as cardiologists do. Thus, we propose a **LaUER**, a **L**anguage-**a**nchored **U**nified **E**CG **R**epresentation Learning framework for above tasks, like cardiologists. To approach the view of cardiologists, the ECG image is introduced to complement the ECG signal. Using clinical text, we design a multi-granularity language-anchored multimodal pre-training strategy to guide ECG representation learning in the language space, including a semantic-granularity learning, a concept-granularity learning, and a token-granularity learning. Through progressive guidance, a unified ECG representation is ultimately obtained that exhibits both detectability and interpretability. To evaluate the performance of LaUER, we conduct linear probing, zero-shot classification, and report generation evaluation on three ECG benchmark datasets, comparing it with the state-of-the-arts (SOTAs). Experimental results show that LaUER outperforms the SOTAs in both linear probing and zero-shot classification, while achieving performance close to SOTAs in report generation at a significantly lower cost compared to SOTAs.

1 INTRODUCTION

Annually, over 300 million electrocardiograms (ECGs) are produced for the purpose of examining cardiovascular diseases (CVDs) (Hannun et al., 2019). Faced with the immense demand for ECG diagnostic services, computer-aided diagnosis model has assumed particular significance. Nevertheless, clinical ECG data are frequently withheld from public disclosure. ECG representations mined based on the conventional supervised learning paradigm struggle to achieve satisfactory results on target-domain data (Wang et al., 2026). It has become crucial to perform CVD (cardiovascular disease) diagnosis without fine-tuning in the target field. Recently, a representation learning paradigm based on ECG-Text alignment has offered insights for this (Li et al., 2023; Liu et al., 2024a;b). However, when cardiologists diagnose CVDs, the diagnostic process includes disease detection results and interpretation of cardiac state. From the viewpoint of language granularity, detection operates at a macroscopic level, whereas interpretation focuses on intricate details. Learning a unified ECG representation for detection and interpretation poses a challenge.

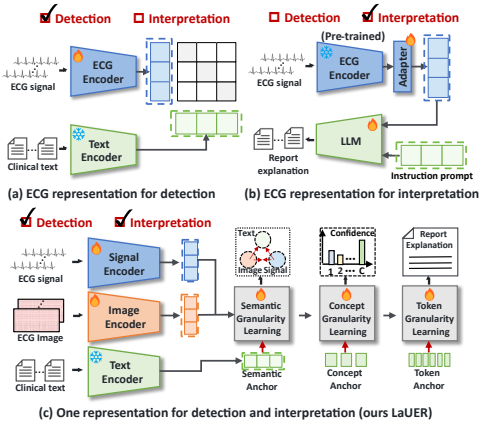

Figure 1: Research illustration.

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.

At present, ECG representation learning has shifted from *single-modal* self-supervision toward *cross-modal* alignment. As shown in Figure 1a, one line of work aligns the ECG signals with clinical reports through contrastive learning, producing a discriminative representation for zero-shot classification (Li et al., 2023; Gao et al., 2024; Liu et al., 2024a; Wang et al., 2025; Niu et al., 2026). Another line extends a pretrained text-only language model to the ECG modality (in Figure 1b), by projecting ECG representation into language embedding space and instruction-tuning the model to narrate reports (Liu et al., 2024b; Zhao et al., 2025; Wan et al., 2025; Lan et al., 2025; Xie et al., 2025). However, the representations learned by both lines can only be associated with specific language granularity. The classification-oriented approach focuses on the discriminability of diseases, hence the ECG representation tends to be more aligned with coarse-grained concepts of CVDs. The generation-oriented approach emphasizes the interpretability of diseases, which focuses on fine-grained linguistic expression. We contend that when a cardiologist detects and interprets CVDs, the fundamental knowledge they rely on is unified, differing only in the granularity of linguistic expression. ***How to enable ECG representations to associate language embeddings of different granularities?*** This question is the key to our research.

To tackle the above issue, we propose a **LaUER**, a **L**anguage-**a**nchored **U**nified **E**CG **R**epresentation Learning framework. ECG signals reflect the physiological electrical processes of the heart, while ECG images are the actual format that cardiologists read. The two modalities are used as the ECG data in LaUER. A multi-granularity language-anchored multimodal pre-training strategy is designed (in Figure 1c), including *semantic-granularity* learning (SGL), *concept-granularity* learning (CGL), and *token-granularity* learning (TGL). In SGL, clinical text is treated as semantic anchors to facilitate mapping of signal and image representations onto the global semantic knowledge inherent in clinical language through an asymmetric cross-modal alignment task. In CGL, a standardized CVD query bank is built that provides concept anchors. After the fusion of signal and image representations, the ECG representation is associated with coarse-grained terms of CVDs by query-guided tracing. In TGL, guided by text token anchors, the autoregressive decoder is employed to predict tokens from the ECG representation, strengthening the association between the ECG representation and fine-grained description of CVD.

In summary, our contributions are as follows:

- **One unified ECG representation for detection and interpretation.** Using ECG signal, image, clinical text, we propose a **LaUER**, a **L**anguage-**a**nchored **U**nified **E**CG **R**epresentation Learning framework, which can, like cardiologists, simultaneously detect diseases and interpret cardiac state.
- **A coarse-to-fine pretraining strategy for ECG representation learning.** We employ a multi-granularity language-anchored multimodal pre-training strategy to guide the relevance from coarse-to-fine between ECG representation and clinical language at semantic-granularity, concept-granularity, and token-granularity.
- **A clinically curated multimodal corpus.** We consolidate disease-label candidates into a canonical vocabulary to annotation MEETI and produce the disease query bank. The entire process calibrates by SNOMED-CT code and sampling review of clinical specialists.
- **State-of-the-art results on standard downstream benchmarks.** Pretrained on MEETI, compared to SOTAs on three ECG benchmark datasets, the performance LaUER is the best in linear probing, zero-shot classification, and near-SOTA in report generation with significantly lower computational cost.

2 Related Work

ECG representation learning for detection. Self-supervised ECG learning uses temporal and patient-level structure to learn transferable features, as in CLOCS (Kiyasseh et al., 2021), TS-TCC (Eldele et al., 2021), and ECGFM (McKeen et al., 2024). ECG–text approaches, including METS (Li et al., 2023), MERL (Liu et al., 2024a), and MELP (Wang et al., 2025), additionally connect physiological signals with diagnostic language. Their contrastive readouts motivate our semantic-granularity stage. A report-level embedding, however, does not explicitly specify how individual disease queries should attend to ECG evidence. SuPreME (Cai et al., 2025) motivates the

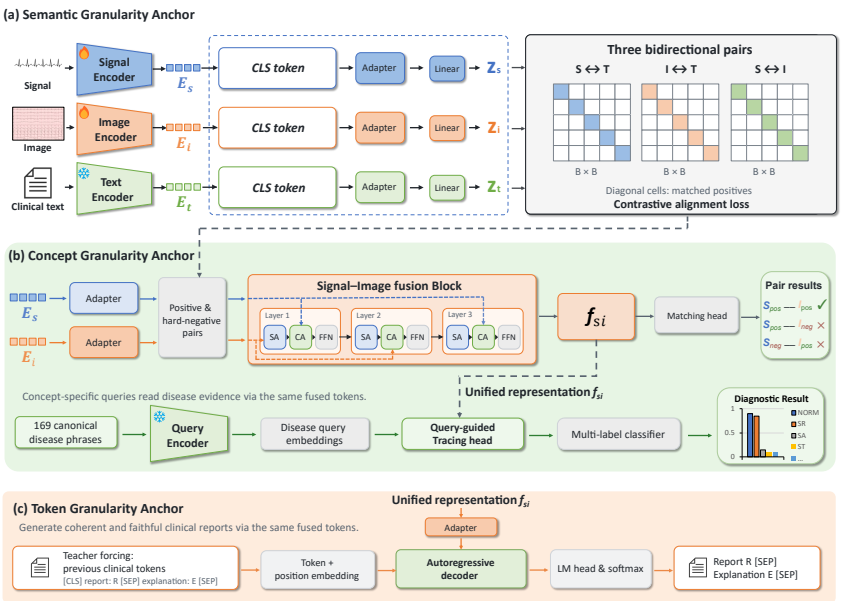

Figure 2: Overview of LaUER. The (a) semantic, (b) concept, and (c) token granularity anchors learn one fused representation $\boldsymbol{F}^{si}$, which is read out both discriminatively (multi-label classification) and generatively (report generation). Frozen modules are marked.

query-guided multi-label readout used in our concept-granularity stage. We combine this readout with signal–image fusion and subsequently train a token-level decoder on the same fused memory.

ECG representation learning for interpretation. PULSE (Liu et al., 2024b), ECG-Chat (Zhao et al., 2025), MEIT (Wan et al., 2025), and GEM (Lan et al., 2025) investigate language-based ECG interpretation. These approaches motivate evaluating not only disease discrimination but also the content of generated descriptions. The distinction between detection and generation is not absolute: CG-DMER (Niu et al., 2026) also couples contrastive and generative learning. Our design focuses on a progressive semantic–concept–token curriculum and a shared signal–image memory rather than claiming that joint discriminative and generative learning is itself unprecedented. Classification results, cardiologist ratings, and qualitative generation diagnostics are reported separately; the case study alone does not establish benchmark-level generation accuracy.

Language-anchored multimodal learning. CLIP (Radford et al., 2021) establishes contrastive alignment across modalities, while ALBEF (Li et al., 2021) and BLIP (Li et al., 2022) motivate aligning representations before fusion and combining understanding with generation. LaUER uses frozen MedCPT (Jin et al., 2023) backbones to supply report-level and concept-level language features. The projection layers remain trainable: a frozen language backbone should not be confused with a completely fixed shared embedding space. The contribution is the coordinated use of three language granularities for ECG, not freezing a pretrained encoder in isolation.

3 Method

LaUER pretrains a shared signal–image representation on MEETI using cumulative semantic-, concept-, and token-granularity objectives (Figure 2). It supports zero-shot classification without target-domain training, linear probes fitted with downstream labels, and autoregressive report generation.

3.1 Multimodal Pre-training Corpus

Our pre-training corpus is built on MEETI (Zhang et al., 2025), a multimodal ECG resource derived from MIMIC-IV-ECG (Gow et al., 2023) that provides, for each of 784,680 ten-second 12-lead

recordings, a raw signal, a plotted image, beat-level quantitative parameters (per-beat intervals and amplitudes), and a detailed interpretation generated by an LLM from the cardiologist-written report and the extracted parameters. We use the first two as our signal and image views and concatenate the report with the interpretation as the language view, i.e.,`report [SEP] interpretation`. We consolidate disease-label candidates into a canonical vocabulary, calibrate their terminology against SNOMED-CT, and have clinical specialists review a sample. This yields $C = 169$ concepts and record-level multi-hot targets for CGL. Downstream disease prompts undergo the same terminology-calibration and sampled-review workflow; they specify candidate classes for inference, not downstream training supervision. Separately, we render a paired image for each signal-only downstream record and conduct sampled clinician review of the images (Appendix A). These resources supply training labels, inference prompts, and input images, respectively.

3.2 Multimodal Encoding

ECG Signal/Image Encoding. The signal and image encoders are separate trainable tiny ViTs, each with 12 blocks, three attention heads, and width $d_{\text{ecg}} = 192$. The signal $\boldsymbol{x}^s \in \mathbb{R}^{12\times T}$ has $T = 5{,}000$ samples per lead. The image $\boldsymbol{x}^i \in \mathbb{R}^{3\times H_{\text{img}}\times W_{\text{img}}}$ has height $H_{\text{img}} = 720$ and width $W_{\text{img}} = 1{,}000$. Each lead is divided into $m = 40$ temporal patches: 125 samples in the signal or 60×25 pixels in the image. To retain lead identity and temporal position, patch (k, j) uses learnable vectors $\boldsymbol{e}_k, \boldsymbol{p}_j \in \mathbb{R}^{d_{\text{ecg}}/2}$:

$$\boldsymbol{h}_{k,j}^{(0)} = \text{PatchEmbed}(x_{k,j}) + W_p[\boldsymbol{e}_k; \boldsymbol{p}_j] + \boldsymbol{b}_p, \tag{1}$$

where $[\cdot\,;\,\cdot]$ concatenates the two position vectors and $W_p, \boldsymbol{b}_p$ define a single affine projection. A prepended CLS token gives $M = 12m + 1 = 481$ tokens. The encoders output $E^s, E^i \in \mathbb{R}^{M\times d_{\text{ecg}}}$, with corresponding patches indexing the same lead and time segment.

Clinical Text Encoding. MedCPT (Jin et al., 2023) supplies two frozen language encoders. Its Article Encoder maps `report [SEP] interpretation` to $E^t \in \mathbb{R}^{N_{\text{text}}\times d_{\text{text}}}$, where $d_{\text{text}} = 768$ and N_{text} is the text-token count. Its CLS feature anchors SGL. The Query Encoder maps the canonical disease phrases to cached CLS features $Q \in \mathbb{R}^{C\times d_{\text{text}}}$ for CGL. ECG/text encoder details are provided in Appendix B.

3.3 Multi-granularity Language-anchored Multimodal Pre-training

Semantic Granularity Learning (SGL): Modality-pair-weighted Contrastive Alignment. Each modality's CLS feature passes through a separate SGL adapter to an ℓ_2-normalized embedding $\boldsymbol{z} \in \mathbb{R}^{d_{\text{con}}}$, with $d_{\text{con}} = 512$. We align signal, image, and clinical text using bidirectional InfoNCE (van den Oord et al., 2018):

$$\mathcal{L}_{a\to b} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\big(s(\boldsymbol{z}_i^a, \boldsymbol{z}_i^b)/\tau\big)}{\sum_{j=1}^{B}\exp\big(s(\boldsymbol{z}_i^a, \boldsymbol{z}_j^b)/\tau\big)}, \quad \mathcal{L}_{ab} = \tfrac{(\mathcal{L}_{a\to b}+\mathcal{L}_{b\to a})}{2}, \quad a, b \in \{s, i, t\} \tag{2}$$

where B is the batch size, $s(\,\cdot\,,\,\cdot\,)$ is dot-product similarity between normalized embeddings, and τ is a learnable temperature.

The two directions within each modality pair are weighted equally. The asymmetry lies instead in the weights *between pairs*: we assign $\alpha_{st} = \alpha_{it} = 0.4$ to the ECG–text pairs and $\alpha_{si} = 0.2$ to signal–image alignment. This emphasizes the two language anchors rather than treating alignment between two views of the same recording as equally important. The SGL objective is

$$\mathcal{L}_{\text{SGL}} = \alpha_{st}\,\mathcal{L}_{st} + \alpha_{it}\,\mathcal{L}_{it} + \alpha_{si}\,\mathcal{L}_{si}. \tag{3}$$

Concept Granularity Learning (CGL): Query-Guided Tracing of Disease Evidence. LaUER follows a "fuse after align" design (Li et al., 2021). Separate CGL adapters map all signal and image tokens to $S, I \in \mathbb{R}^{M\times d_{\text{fusion}}}$, with $d_{\text{fusion}} = 768$. Three fusion blocks alternate cross-attention to the original projected memories, beginning with $F^{si(0)} = I$:

$$F^{si(l)} = \text{FusionBlock}_l(F^{si(l-1)}, M_l), \qquad (M_1, M_2, M_3) = (S, I, S), \tag{4}$$

where each block applies self-attention, cross-attention, and a feed-forward sublayer, yielding $F^{si} = F^{si(3)}$. For every matched signal–image pair, off-diagonal SGL similarities are used to sample one incorrect image and one incorrect signal. These pairs yield memories F^{+}, F^{i-}, and F^{s-}. A two-logit signal–image matching head reads the fused CLS token; its softmax gives the match probability $p_\theta(F^{si})$. The matching loss is

$$\mathcal{L}_{\text{match}} = -\frac{1}{3B}\sum_{b=1}^{B}\left[\log p_\theta(F_b^{+}) + \log(1 - p_\theta(F_b^{i-})) + \log(1 - p_\theta(F_b^{s-}))\right]. \tag{5}$$

The *Query-guided tracing head*, we uses the cached disease bank Q to read evidence from F^{si}. A trainable affine projection and LayerNorm adapt these frozen language features; three Transformer blocks then apply query self-attention, cross-attention to F^{si}, and a feed-forward sublayer. For record b,

$$\begin{aligned} Q^{(0)} &= \mathrm{LN}(QW_q + \mathbf{1}\boldsymbol{b}_q^{\top}), \\ Q_b^{(l)} &= \mathrm{TraceBlock}_l(Q_b^{(l-1)}, F_b), \quad l = 1, 2, 3, \\ \boldsymbol{h}_{b,c} &= [\mathrm{LN}(Q_b^{(3)})]_c, \qquad \ell_{b,c} = \boldsymbol{w}_{\text{clf}}^{\top}\boldsymbol{h}_{b,c} + b_{\text{clf}}. \end{aligned} \tag{6}$$

The initial query set is shared across records: $Q_b^{(0)} = Q^{(0)}$. Here $W_q \in \mathbb{R}^{d_{\text{text}} \times d_{\text{fusion}}}$ and $\boldsymbol{b}_q \in \mathbb{R}^{d_{\text{fusion}}}$ are trainable. A shared affine *multi-label classifier*, with $\boldsymbol{w}_{\text{clf}} \in \mathbb{R}^{d_{\text{fusion}}}$, produces one logit per query. Predictions use independent sigmoid probabilities $\sigma(\ell_{b,c})$, not a softmax across diseases. For multi-hot targets $y_{b,c} \in \{0, 1\}$, the loss is

$$\mathcal{L}_{\text{clf}} = -\frac{1}{BC}\sum_{b=1}^{B}\sum_{c=1}^{C}\left[y_{b,c}\log\sigma(\ell_{b,c}) + (1 - y_{b,c})\log(1 - \sigma(\ell_{b,c}))\right]. \tag{7}$$

Query-conditioned cross-attention links each disease score to ECG tokens; these attention patterns are not, by themselves, validated clinical explanations. CGL combines signal–image matching and multi-label binary cross-entropy:

$$\mathcal{L}_{\text{CGL}} = \beta\,\mathcal{L}_{\text{match}} + (1 - \beta)\,\mathcal{L}_{\text{clf}}, \tag{8}$$

where β controls the relative contribution of the two supervisions.

Token Granularity Learning (TGL): Autoregressive Generation. TGL trains an autoregressive text decoder to generate the report and interpretation from the same fused ECG memory. A two-layer MLP projects F to the decoder width, and a six-layer Transformer reads it through cross-attention. Causal teacher forcing predicts the continuous target stream `[CLS] report:` R `[SEP] explanation:` E `[SEP]`. The first separator ends the report; the second ends generation. One objective covers the non-padding target positions $\mathcal{T}$ in a minibatch:

$$\mathcal{L}_{\text{TGL}} = -\sum_{(b,k)\in\mathcal{T}} \omega_{b,k}\log p_\theta(w_{b,k} \mid w_{b,<k}, F_b), \tag{9}$$

where $\omega_{b,k}$ normalizes target-region lengths so that the longer interpretation does not dominate the report (Appendix B). At inference, generation begins with `[CLS] report:` and stops at the second separator. Token supervision encourages retention of details expressed in the reference text; it does not guarantee faithful recovery of every numerical finding.

Training is cumulative: SGL → SGL+CGL → SGL+CGL+TGL. Each stage initializes from the preceding checkpoint and retains its objectives. Thus $\mathcal{L} = \mathcal{L}_{\text{SGL}} + \mathcal{L}_{\text{CGL}} + \mathcal{L}_{\text{TGL}}$ is the *final-stage* objective, rather than the loss used from the start of training.

3.4 Detection and Interpretation for Target-domain Data

The task of disease detection and interpretation within the target domain fundamentally constitutes zero-shot classification and report generation. For zero-shot classification, although SGL can perform this task based on semantic anchors through a similarity matrix, CGL with its query-guided tracing head based on concept anchors, achieves better prediction results. In report generation, TGL utilizes a trained lightweight autoregressive decoder to generate clinical report. None of the aforementioned tests requires fine-tuning on the target-domain data.

4 EXPERIMENTS

4.1 PRETRAINING SETUP

LaUER is pretrained on 784,680 MEETI recordings with a 96/2/2 training/validation/test split (Section 3.1), without data augmentation. The cumulative SGL → SGL+CGL → SGL+CGL+TGL schedule introduces semantic alignment, 169-concept supervision, and text generation in turn. Training uses two NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, BF16, and AdamW. Per-GPU batch sizes are 468/108/64; two accumulation steps give effective batches of 1,872/432/256. The epoch budgets are 50/5/5, with validation-based checkpoint selection and early stopping. For the three complete training sequences (seeds 42, 66, and seed 20250915), SGL uses learning rate 5.8×10^{-4} for the ECG encoders and semantic projections. CGL uses 2×10^{-5}/10^{-6} for these components and 10^{-4} for fusion and query modules. TGL uses 10^{-5}/10^{-7}/8×10^{-7} for encoders/projections/fusion and 10^{-4} for the decoder. Warmup precedes cosine/constant/cosine schedules, respectively. MedCPT remains frozen; all other components continue training. Appendix C.1 gives the remaining optimizer and selection settings.

4.2 DOWNSTREAM TASKS AND EVALUATION

We evaluate four PTB-XL tasks (Wagner et al., 2020), CPSC2018 (Liu et al., 2018), and CSN (Zheng et al., 2020; 2022) using MELP's split files (Wang et al., 2025): official PTB-XL folds 1–8/9/10 and fixed CPSC2018/CSN partitions. Only records with task-relevant labels are retained (Table 1); image construction is described in Appendix A. We report macro-AUROC (%) over classes with both outcomes, weighting tasks equally. Full-model results average pretraining seeds 42 and 66. Literature rows and reported averages are external references, not harmonized reruns; their numerical ranks are descriptive.

Table 1: Downstream cohorts. Label fractions subsample training records only.

Task	Classes	Train	Validation	Test
PTB-XL Rhythm	12	16,832	2,100	2,098
PTB-XL Sub	23	17,084	2,146	2,158
PTB-XL Form	19	7,197	901	880
PTB-XL Super	5	17,084	2,146	2,158
CPSC-2018	9	4,950	551	1,376
CSN	38	16,546	1,860	4,620

4.3 LINEAR EVALUATION

Linear evaluation tests whether disease information can be read out without updating the pretrained encoders. We fit separate affine classifiers to the frozen signal, image, and fused CLS features of the final SGL+CGL+TGL model, using 1%, 10%, and 100% of downstream training labels. Table 2 compares the three readouts with published baselines; probe initialization and subsampling use seed 42 (Appendix C.2).

Table 2: Linear-evaluation AUROC (%). Baselines: MELP Table 2 (Wang et al., 2025); ours: two-seed final-stage means. Bold/underline: highest/second-highest. SDs: Appendix E.1.

	Rhythm			Subclass			Form			Superclass			CPSC2018			CSN		
Method	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
SimCLR (2020)	51.41	69.44	77.73	60.84	68.27	73.39	54.98	56.97	62.52	63.41	69.77	73.53	59.78	68.52	76.54	59.02	67.26	73.20
BYOL (2020)	41.99	74.40	77.17	57.16	67.44	71.64	48.73	61.63	70.82	71.70	73.83	76.45	60.88	74.42	78.75	54.20	71.92	74.69
BarlowTwins (2021)	50.12	73.54	77.62	62.57	70.84	74.34	52.12	60.39	66.14	72.87	75.96	78.41	55.12	72.75	78.39	60.72	71.64	77.43
MoCo-v3 (2021)	51.38	71.66	74.33	55.88	69.21	76.69	50.32	63.71	71.31	73.19	76.65	78.26	62.13	76.74	75.29	54.61	74.26	77.68
SimSiam (2021)	49.30	69.47	75.92	62.52	69.31	76.38	55.16	62.91	71.31	73.15	72.70	75.63	58.35	72.89	75.31	58.25	68.61	77.41
TS-TCC (2021)	43.34	69.48	78.23	53.54	66.98	77.87	48.04	61.79	71.18	70.73	75.88	78.91	57.07	73.62	78.72	55.26	68.48	76.79
CLOCS (2021)	47.19	71.88	76.31	57.94	72.55	76.24	51.97	57.79	72.65	68.94	73.36	76.31	59.59	77.78	77.49	54.38	71.93	76.13
W2V2+CMSC+RLM (2022)	76.24	86.34	92.05	69.10	80.71	85.01	52.72	67.81	80.72	81.15	84.88	85.53	75.70	88.16	92.61	65.65	78.82	87.87
ASTCL (2024)	52.38	71.98	76.05	61.86	68.77	76.51	44.14	60.93	66.99	72.51	77.31	81.02	57.90	77.01	79.51	56.40	70.87	75.79
CRT (2023)	47.44	73.52	74.41	61.98	70.82	78.67	46.41	59.49	68.73	69.68	78.24	77.24	58.01	76.43	82.03	56.21	73.70	78.80
ECGFM (2024)	81.45	91.59	92.70	73.24	81.91	86.07	60.95	74.99	85.54	78.67	84.80	86.47	82.18	89.52	93.26	71.51	83.17	88.89
ST-MEM (2024)	51.12	65.44	74.85	54.12	57.86	63.59	55.71	59.99	66.07	61.12	66.87	71.36	56.69	63.32	70.39	59.77	66.87	71.36
HeartLang (2025)	62.08	76.22	90.34	64.68	79.34	88.91	58.70	63.99	80.23	78.94	85.59	87.52	60.44	66.26	77.87	57.94	68.93	82.49
MERL (2024a)	53.33	82.88	88.34	64.90	80.56	84.72	58.26	72.43	79.65	82.39	86.27	88.67	70.33	85.32	90.57	66.60	82.74	87.95
MELP (2025)	**88.83**	**94.65**	**96.91**	79.22	84.40	87.46	63.41	76.71	83.30	85.82	87.61	87.87	**88.54**	91.75	94.32	78.25	84.83	90.17
Ours–Signal	77.03	88.77	92.02	75.13	87.08	91.87	59.69	75.10	83.92	84.09	88.19	90.08	77.96	91.39	94.44	73.19	88.90	94.72
Ours–Image	79.60	91.33	93.41	80.46	89.67	92.93	58.34	74.76	86.11	86.57	**90.74**	91.81	83.07	93.82	95.89	75.52	88.79	94.67
Ours–Fusion	86.38	92.72	93.73	**83.39**	**90.30**	**93.23**	**64.58**	**80.63**	**86.53**	**87.30**	90.56	**92.06**	85.54	**93.92**	**95.97**	**78.36**	**90.85**	**95.34**

Table 3: Cross-dataset AUROC (%); baselines: MELP Table 4 (Wang et al., 2025). P/C/S: PTB-XL Superclass/CPSC2018/CSN; ratio: source-label budget. Baseline averages are as published.

Method	Zero-shot	Ratio	P→C	P→S	C→P	C→S	S→P	S→C	Avg.
Literature: source-domain linear evaluation									
SimCLR (2020)	No	100%	69.62	73.05	56.65	66.36	59.74	62.11	65.22
BYOL (2020)	No	100%	70.27	74.01	57.32	67.56	60.39	63.24	65.63
BarlowTwins (2021)	No	100%	68.98	72.85	55.97	65.89	58.76	61.35	64.13
MoCo-v3 (2021)	No	100%	69.41	73.29	56.54	66.12	59.82	62.07	64.21
SimSiam (2021)	No	100%	70.06	73.92	57.21	67.48	60.23	63.09	65.33
TS-TCC (2021)	No	100%	71.32	75.16	58.47	68.34	61.55	64.48	66.55
CLOCS (2021)	No	100%	68.79	72.64	55.86	65.73	58.69	61.27	63.83
ASTCL (2024)	No	100%	69.23	73.18	56.61	66.27	59.74	62.12	64.19
CRT (2023)	No	100%	70.15	74.08	57.39	67.62	60.48	63.33	65.51
ST-MEM (2024)	No	100%	76.12	84.5	62.27	75.19	73.05	64.66	72.63
MERL (2024a)	Yes	0%	**88.21**	78.01	76.77	76.56	74.15	**82.86**	79.42
MELP (2025)	Yes	0%	87.75	74.11	77.89	80.32	74.67	82.72	79.58
Ours	Yes	0%	75.50	**92.08**	**78.15**	**86.95**	**79.79**	82.44	**82.48**

Table 4: Native-label zero-shot AUROC (%). Baselines: MELP Table 3 (Wang et al., 2025); gains are percentage-point differences from its published MELP row.

Method	CSN	Rhythm	Form	Subclass	Superclass	CPSC2018	Avg.
MERL (2024a)	74.4	78.5	65.9	75.7	74.2	82.8	75.3
MELP (2025)	77.6	85.4	69.1	81.2	76.2	84.2	79.0
Ours	83.19	88.70	66.77	83.05	85.68	88.94	82.72
Gain vs. MELP (pp)	**+5.59**	**+3.30**	**-2.33**	**+1.85**	**+9.48**	**+4.74**	**+3.72**

Fusion exceeds both unimodal readouts in 17 of 18 task–label-budget settings and leads 13 displayed columns. At the 1% budget, its Form AUROC is 64.58, compared with 59.69 for signal and 58.34 for image features. Its gains over image features narrow with more labels on Rhythm and Form; image remains higher on Superclass at 10% (90.74 versus 90.56). These results favor the fused readout, although the unequal feature dimensions (192 versus 768) do not isolate modality complementarity from capacity.

4.4 Zero-shot Classification

Both zero-shot experiments use the *query-conditioned* readout and the fixed four-component prompt ensemble in Appendix C.3, without training or fine-tuning on downstream records. In native-label evaluation, disease prompts correspond to each benchmark's own label vocabulary (Table 4).

LaUER reaches 82.72% mean AUROC, 3.72 percentage points above the published MELP mean. It exceeds MELP on five of six tasks: the largest gain is 9.48 points on Superclass, whereas Form remains lower by 2.33 points (66.77 versus 69.10). The improvement is therefore task-dependent rather than uniform.

The cross-dataset experiment maps a source concept vocabulary to target labels and excludes records with unmapped positive labels (Table 3). LaUER's arrows denote these vocabulary mappings, not source-domain training; source-label budgets apply only to the published source-trained baselines.Cross-dataset AUROC averages 82.48%, ranging from 75.50 for P→C to 92.08 for P→S. Because the mapped concepts and retained cohorts differ by direction, these scores do not compare the same classification problem. Historical test results informed prompt-preset selection, so they are not untouched-test estimates.

4.5 Report Generation

Six cardiologists each evaluated generated text for all 200 randomly sampled MEETI test ECGs using GEM's seven criteria (Lan et al., 2025) (Table 5; protocol in Appendix C.4). Published GEM scores concern different cases and eight cardiologists, so they are external references rather than paired comparisons.Accuracy (4.5) exceeds completeness (4.2), distinguishing correct findings from

Table 5: Cardiologist ratings (mean ± SD on a five-point scale).

Model	Analytical Relevance	Analytical Accuracy	Analytical Completeness	Reasoning Quality	Findings Novelty	Clinical Value	Overall Satisfaction
GEM (7B)	4.6 ± 0.60	4.4 ± 0.80	4.6 ± 0.57	4.6 ± 0.64	3.9 ± 1.25	4.3 ± 0.89	4.4 ± 0.82
Ours (267.6M)	4.3 ± 0.52	4.5 ± 0.48	4.2 ± 0.60	4.5 ± 0.67	4.7 ± 1.03	4.0 ± 0.85	4.3 ± 0.64

Table 6: Stage-wise AUROC (%). SGL/CGL/TGL average seeds 42, 66, and 20260915 ($n = 3$). Probe sampling uses seed 42.

	Zero-shot		Linear evaluation (1%)		
Stage	Semantic-based	Concept-based	Signal	Image	Fusion
SGL	66.17	N/A	74.68	77.38	N/A
SGL+CGL	66.13	82.94	74.38	77.26	80.49
SGL+CGL+TGL	66.11	82.72	74.52	77.26	80.92

Table 7: Seed-42 supervision and prompt comparisons (mean AUROC, %). Panel (a) uses CGL; panel (b) fixes the TGL checkpoint.

(a) Supervision/model configuration

		Zero-shot		Linear (1%)		
Scheme	Labels	Semantic-based	Concept-based	Signal	Image	Fusion
SuPreME-derived	239	65.66	76.45	68.21	72.34	77.61
Optimized	169	66.17	83.20	75.55	76.54	80.64

(b) Prompt formulation

	Query-conditioned	
Prompt	Native	Cross-domain
CKEPE	79.34	80.33
Our prompts	81.87	81.59

comprehensive coverage. Likewise, high novelty (4.7) does not imply clinical value (4.0). Neither the small score gaps nor the parameter counts establish equivalence or matched computational efficiency with GEM.

4.6 Ablation Studies

We ablate training stages, supervision labels, disease prompts, and SGL loss weights using zero-shot classification and 1%-label linear probes. AUROC is averaged equally across six tasks (Appendix C.5). Contrastive-similarity scoring is available throughout; query-conditioned scoring requires CGL.

After TGL, query-conditioned scoring changes from 82.94 to 82.72, while fused linear evaluation changes from 80.49 to 80.92 (Table 6). The unequal seed sets and additional optimization prevent attributing these differences solely to TGL.

In Table 7(a), the 169-label configuration gains 6.75/3.03 points in query-conditioned/fused-probe AUROC over the 239-label configuration with fixed prompts. Pooling changes from lead-attention/CLS-plus-attention to CLS, and selected epochs also differ, so this comparison does not isolate label curation. Panel (b) fixes checkpoint `3_v4.pt` and uses task-only prompts without ensembling. Our prompts gain 2.53/1.26 points in native/cross-domain AUROC; the latter retains mapped-positive records, unlike Table 3. Terminology calibration and clinician review are not independently ablated (Appendix C.5).

Table 8: SGL-weight sensitivity at SGL (seed 42), with fixed semantic-based zero-shot prompts and 1%-label probes. Scores are six-task mean AUROCs (%).

				Linear evaluation (1%)	
α_{st}	α_{it}	α_{si}	Semantic-based zero-shot	Signal	Image
0.30	0.30	0.40	65.51	76.46	77.90
0.35	0.35	0.30	65.91	75.55	76.65
0.40	0.40	0.20	66.18	75.92	76.53
1.00	1.00	1.00	65.67	76.02	77.67

Among the three unit-sum settings in Table 8, $(0.40, 0.40, 0.20)$ has the highest contrastive-similarity score (66.18), whereas $(0.30, 0.30, 0.40)$ favors both probes. No setting dominates all readouts; $(1, 1, 1)$ also changes the loss scale.

4.7 Representation Analysis

Figure 3 contrasts raw features (L2 normalization/PCA50) with nine-class Ridge scores and BCE logits for 2,464 single-label CPSC2018 records at two seed-42 checkpoints. These downstream readouts use training labels, with Ridge regularization selected on validation AUROC; they are distinct from the pretrained Query-guided tracing head (Appendix C.6).

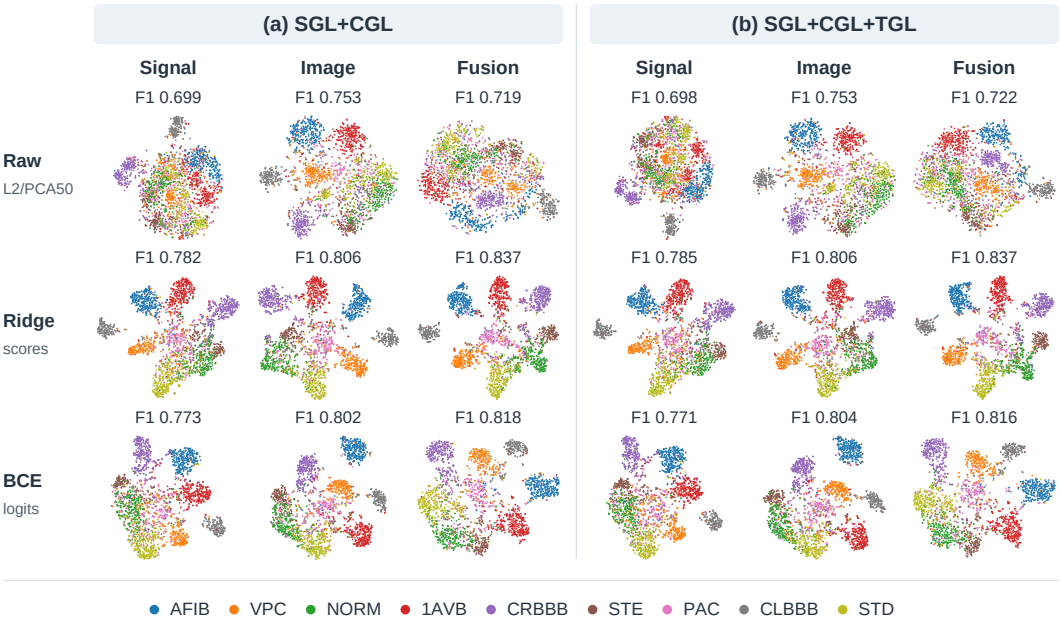

Figure 3: CPSC2018 representations, with fixed records and class colors. Scores: pre-t-SNE 20-NN macro-F1 (five-fold CV). Each panel is fitted independently; coordinates are not comparable.

Without supervised readout, fusion's neighborhood macro-F1 falls between signal and image at both stages. After TGL, the scores are 0.698, 0.753, and 0.722 for signal, image, and fusion. Ridge readout changes this ordering to 0.785, 0.806, and 0.837; BCE readout gives 0.771, 0.804, and 0.816. Fusion thus ranks highest under both readouts, with the same ordering already present before TGL. This does not mean every projection metric favors fusion: after TGL, its BCE-space two-dimensional silhouette is 0.271, below image's 0.283. Neither tighter-looking clusters nor a single projection score fully describes the readout advantage. Since $z = Wx + b$ induces $\|z_i - z_j\|_2^2 = (x_i - x_j)^\top W^\top W(x_i - x_j)$, this is consistent with task-related reweighting, not tighter raw clusters. These are descriptive diagnostics: the cohort includes readout-training records and readouts are not refitted within kNN folds. Appendix D adds CSN views.

5 Discussion and Conclusion

We propose a LaUER, a language-anchored unified ECG representation learning framework via ECG signal, image, and clinical text for disease detection and interpretation. LaUER uses semantic, concept, and token anchors to perform a multi-granularity language-anchored multimodal pre-training by contrastive alignment, query-guided tracing, and autoregressive generation. A clinically curated multimodal corpus is designed to provide disease query bank for MEETI. The learned ECG representation supports concept-based zero-shot classification and autoregressive report generation without fine-tuning on target-domain data. Extensive experiments are presented on three ECG benchmark datasets. The experiments show that LaUER outperforms the SOTAs in terms of linear probing and zero-shot classification. Meanwhile, in terms of report generation, LaUER achieves near-SOTA performance with significantly lower computational overhead.

REFERENCES

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*, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9640–9649, 2021.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

Yuan Gao, Sangwook Kim, David E. Austin, and Chris McIntosh. MEDBind: Unifying language and multimodal medical data embeddings. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Joachim Greenbaum, Payam Eslami, Adam Sana, Leo Celi, and Roger Mark. MIMIC-IV-ECG diagnostic electrocardiogram matched subset (version 0.1.0). PhysioNet, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 21271–21284, 2020.

A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Med.*, 25:65–69, Jan 2019. doi: 10.1038/s41591-018-0268-3.

Jingdong Jin, Hui Wang, Haotian Li, Jing Li, Jimin Pan, and Shenda Hong. Reading your heart: Learning ECG words and sentences via pre-training ECG language model. In *International Conference on Learning Representations (ICLR)*, 2025.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11), 2023.

Dani Kiyasseh, Tingting Zhu, and David A Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *International Conference on Machine Learning (ICML)*, pp. 5606–5615, 2021.

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

Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. Frozen language model helps ECG zero-shot learning. *arXiv preprint arXiv:2303.12311*, 2023.

Junnan Li, Ramprasaath R Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven C H Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pp. 12888–12900, 2022.

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 (ICML)*, 2024a.

Feng Liu, Chengyu Liu, Lina Zhao, Xiaoyang Zhang, Xulin Wu, Xingyuan Xu, Yutong Liu, Cun Ma, Shousong Wei, Zheng He, Junping Li, et al. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology analysis. *Scientific Reports*, 8:12824, 2018.

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

Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*, 2024.

Yujin Na, Mooyoung Park, Yunyang Tae, and Sunghoon Joo. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *arXiv preprint arXiv:2402.09450*, 2024.

Ziwei Niu, Hao Sun, Shujian Bian, Xihong Yang, Lanfen Lin, Yuxin Liu, and Yueming Jin. CG-DMER: Hybrid contrastive-generative framework for disentangled multimodal ECG representation learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Jungwoo Oh, Hyunseung Chung, Joon-myoung Kwon, Dong-gyun Hong, and Eunho Choi. Lead-agnostic self-supervised learning for local and global representations of electrocardiogram. In *Conference on Health, Inference, and Learning (CHIL)*, pp. 338–353, 2022.

PKUDigitalHealth. ecg-plot: a python package for plotting electrocardiogram. `https://github.com/PKUDigitalHealth/ecg_plot`, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

A"aron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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 for deep learning. *Scientific Data*, 7(1):154, 2020.

Zhongwei Wan, Che Liu, Xin Wang, Chaofan Tao, Haowei 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*, 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 (ICML)*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025.

Ning Wang, Peng Feng, Zhengang Ge, Yong Zhou, Bin Zhou, and Zhiyong Wang. Adversarial spatiotemporal contrastive learning for electrocardiogram signals. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):13845–13859, 2024.

Ning Wang, Haiyan Wang, Jian Tan, Panpan Feng, Shihua Li, Zongmin Wang, and Bing Zhou. Ecg-text multi-modal learning for zero-shot detection via time-frequency alignment and medical prompt learning. *Expert Syst. Appl.*, 307:131064, 2026.

Yihan Xie, Sijing Li, Tianwei Lin, Zhuonan Wang, Chenglin Yang, Yu Zhong, Wenjie Yan, Wenqiao Zhang, Xiaogang Guo, Jun Xiao, Yueting Zhuang, and Beng Chin Ooi. HeartcareGPT: A unified multimodal ECG suite for dual signal-image modeling and understanding. *arXiv preprint arXiv:2506.05831*, 2025.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning (ICML)*, pp. 12310–12320, 2021.

Deyun Zhang, Xiang Lan, Shijia Geng, Qinghao Zhao, Sumei Fan, Mengling Feng, and Shenda Hong. MEETI: A multimodal ECG dataset from MIMIC-IV-ECG with signals, images, features and interpretations. *arXiv preprint arXiv:2507.15255*, 2025.

Wei Zhang, Lin Yang, Shijia Geng, and Shenda Hong. Self-supervised time series representation learning via cross reconstruction transformer. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Yubao Zhao et al. ECG-Chat: A large ECG-language model for cardiac disease diagnosis. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 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*, 10:2898, 2020.

Jianwei Zheng, Huimin Guo, and Huimin Chu. A large scale 12-lead electrocardiogram database for arrhythmia study (version 1.0.0). PhysioNet, 2022.