PLARE: FEATURE REPLACEMENT WITH A MIXTURE OF EXPERTS FOR SELF-SUPERVISED ECG REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-supervised models learn electrocardiogram (ECG) representations from unlabeled recordings. ECG waveform shapes vary within and across recordings: an isolated beat may differ from its neighbors, whereas a distinctive shape may recur across beats. This variation motivates selecting feature-processing parameters according to the content of each short waveform segment (patch). To this end, we introduce PLARE (Patch-Level Adaptive Replacement with Experts), a self-supervised ECG model with a patch-level mixture-of-experts (MoE) module. Content-dependent routing (CDR) selects experts based on each patch's convolutional features, and patch feature replacement uses their outputs in place of the original features before token projection. A temporal Transformer then models relationships among the resulting patch tokens to produce a record-level ECG representation. Across four PTB-XL label sets and CPSC2018, PLARE achieved higher AUROC and AUPRC than the evaluated baselines on four of the five tasks under both linear probing and fine-tuning. In ablations, feature replacement achieved a higher average of AUROC and AUPRC than residual addition on four of five tasks under each protocol. Content-dependent routing improved this average over fixed random expert assignment on all five tasks under both protocols. Inference costs were lower than those of larger baselines. Using INCART's beat-level annotations, we found that expert selection distributions in the frozen pretrained model differed between contexts surrounding normal beats and premature ventricular contractions. These findings support content-dependent feature replacement for ECG representation learning. Code is available at `https://anonymous.4open.science/r/PLARE-5FC6`.

1 INTRODUCTION

Self-supervised learning (SSL) trains ECG encoders to map raw recordings to ECG representations using unlabeled data. Patch-based Transformer approaches divide ECG waveforms into short, fixed-length temporal segments, called patches. Shared patch feature processing maps each patch to a token using the same learned parameters across patches, and a temporal Transformer models relationships among the resulting tokens across the recording (Na et al., 2024; Song et al., 2024). Pretraining objectives include masked waveform reconstruction (Na et al., 2024), contrastive learning (Kiyasseh et al., 2021; Wang et al., 2023), and their combination (Song et al., 2024).

Local ECG waveform shapes vary within and across recordings. For illustration, Figure 1 treats each heartbeat as a patch (in practice, PLARE uses shorter, fixed-length waveform segments). An isolated premature ventricular contraction (PVC) produces a patch whose waveform differs from those of neighboring patches (left), whereas the broad QRS morphology of persistent left bundle-branch block (LBBB) recurs across patches (right) (Al-Khatib et al., 2018; Kusumoto et al., 2019). These examples motivate examining whether patch feature construction benefits from content-dependent parameter selection before temporal contextualization.

Shared patch feature processing produces input-dependent features while using the same parameters for every patch. A mixture-of-experts (MoE) module contains multiple learnable networks, called experts, and a router that selects which experts process each input. Patches can thus use and update different expert parameters according to their content. We therefore investigate the utility of content-dependent patch routing and how expert outputs are used for subsequent temporal contextualization.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

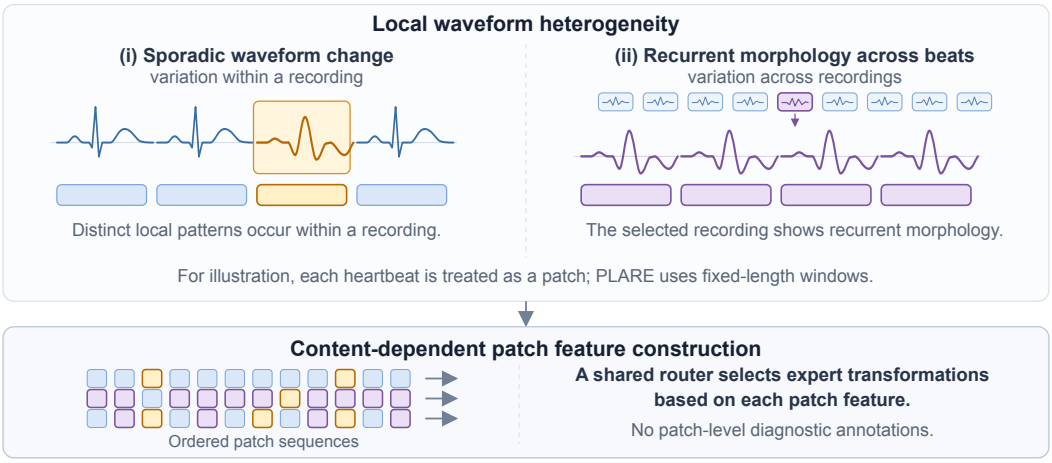

Figure 1: Local waveform heterogeneity: a sporadic waveform change within a recording and recurrent morphology across beats. Beat-aligned patches are used for conceptual illustration; PLARE uses fixed-length temporal windows. Colors indicate schematic waveform content. Expert selection is conditioned on local patch features.

To this end, we propose PLARE, a self-supervised ECG model that adapts local processing to waveform content through a patch-level MoE module. Content-dependent routing (CDR) selects experts based on each patch's features, allowing patches to use different processing parameters rather than relying only on a shared transformation. Patch feature replacement uses the selected experts' outputs as the patch features for temporal contextualization, rather than adding them to the original patch features. A temporal Transformer then contextualizes the projected patch tokens to produce a record-level ECG representation. We train the model jointly with masked reconstruction, contrastive learning, and routing regularization.

Across four PTB-XL label sets and CPSC2018, PLARE achieved higher AUROC and AUPRC than the evaluated baselines on four of five tasks under both linear probing and fine-tuning. Controlled ablations assessed both design choices using the mean of AUROC and AUPRC. Under each protocol, residual addition in place of feature replacement lowered this mean on four of five tasks, while fixed random expert assignment in place of content-dependent routing lowered it on all five. PLARE also used fewer parameters and FLOPs and had lower inference latency and peak memory usage than larger baselines at their respective input configurations.

We further examined local routing on INCART, whose beat location and type annotations enable comparisons around normal and ventricular ectopic beats within 12-lead recordings (Yakushenko, 2008). With the pretrained encoder frozen, router selections spanned all four experts, and their distributions differed between the two beat contexts.

These results support constructing local ECG patch features through content-dependent routing and feature replacement before temporal contextualization.

Our contributions are threefold:

- We introduce PLARE, which replaces local convolutional patch features with routed expert outputs before temporal contextualization for self-supervised ECG representation learning.
- We evaluate downstream performance and analyze the effects of content-dependent routing and feature replacement through controlled comparisons.
- We characterize associations between expert selection and annotated beat contexts in a frozen pretrained encoder, complementing the performance comparisons with an analysis of learned allocation.

2 Related Work

Self-supervised ECG representation learning. ST-MEM linearly embeds per-lead patches and CREMA extracts convolutional features; both use temporal Transformers for contextualization and decoders to reconstruct masked waveforms from visible-patch representations (Na et al., 2024; Song et al., 2024). Other approaches explore multiscale and lead-aware representations (Wang et al., 2025; Jin et al., 2025). Contrastive methods align representations of related inputs (Kiyasseh et al., 2021;

Wang et al., 2023); CREMA also aligns record-level representations. ECG-FM combines masked latent prediction and segment-level contrastive learning (McKeen et al., 2025); ECGFM combines contrastive predictive coding, supervised classification, and diagnostic text generation (Zhang et al., 2025b).

Expert-based conditional computation. Mixture-of-experts (MoE) models route inputs to specialized parameter subsets (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022; Zhou et al., 2022; Puigcerver et al., 2023), with applications to time series (Shi et al., 2025) and supervised ECG prediction or adaptation (Hu et al., 1997; Güler & Übeyli, 2005; Tuler et al., 2024; Mansourian et al., 2025). Within encoders, MoViE uses expert feed-forward networks (FFNs) after attention for prediction from ECGs and electronic health records (Zhang et al., 2025a), while MIRA uses token-wise MoE for heterogeneous medical time-series forecasting (Li et al., 2025a).

MAMMOTH replaces the feature-transformation layer for multiple-instance learning in pathology with multi-head experts, pooling patch embeddings into expert slots and producing compact outputs for aggregation (Shao et al., 2026). The Kairos preprint adapts temporal granularity through expert-based tokenization (Feng et al., 2025).

3 PLARE: Patch-Level Adaptive Replacement with Experts

3.1 Architecture Overview

The PLARE ECG encoder maps raw ECG to a record-level representation through a patch feature extractor, patch-level MoE, token projection, and temporal Transformer (Figure 2). In the patch-level MoE, CDR selects experts whose outputs replace the original patch features. Token projection then produces one token per fixed-width patch for temporal contextualization. This combination of patch-level CDR and feature replacement differs from prior approaches that apply experts within Transformer blocks (Li et al., 2025a), pool patch embeddings into expert slots (Shao et al., 2026), or adapt temporal tokenization (Feng et al., 2025).

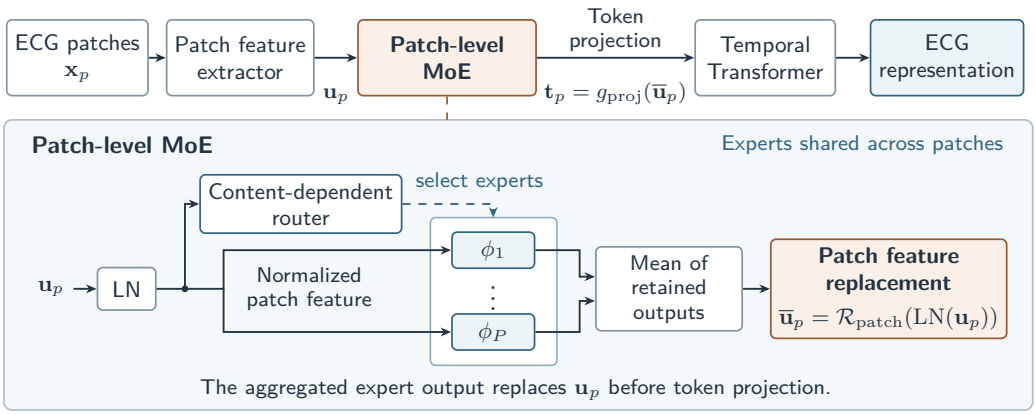

Figure 2: PLARE ECG encoder. **Top:** the patch feature extractor (Conv/Res), patch-level MoE, and token projection construct a patch-token sequence. A temporal Transformer produces the record-level ECG representation. **Bottom:** the patch-level MoE combines content-dependent routing with patch feature replacement. Retained expert outputs are averaged to replace the original patch feature before projection. LN denotes layer normalization. The temporal Transformer retains its standard FFNs and residual connections; the fallback for discarded assignments is defined in Section 3.3.

Following the upper panel of Figure 2, we divide each ECG recording into overlapping temporal patches $\mathbf{x}_p$, each spanning all leads. The shared patch feature extractor produces $\mathbf{u}_p = f_{\text{patch}}(\mathbf{x}_p) \in \mathbb{R}^F$. The patch-level MoE routes these features to selected experts, whose aggregated outputs replace $\mathbf{u}_p$ to give $\overline{\mathbf{u}}_p$. The token projection $g_{\text{proj}} : \mathbb{R}^F \to \mathbb{R}^D$ then produces token $\mathbf{t}_p$. Positional embeddings are added to the tokens, and a learned classification (CLS) token is prepended. A temporal Transformer with N blocks contextualizes this sequence. The final CLS output is the record-level ECG representation used for prediction. Appendix A.2 specifies patch extraction and the full encoder formulation.

3.2 CONTENT-DEPENDENT ROUTING

CDR allows different patches to use different subsets of shared experts, with the choice determined by each patch's convolutional features. The module contains P learned expert networks $\phi_e : \mathbb{R}^F \to \mathbb{R}^F$, indexed by e. Its router computes probabilities from the layer-normalized feature $\mathbf{h}_p^{\text{patch}} = \text{LN}_{\text{patch,in}}(\mathbf{u}_p)$ and selects the k_P highest-scoring experts, where $1 \le k_P \le P$. Training uses noisy Top-k selection for exploration. Capacity filtering retains each expert's highest-scoring assignments up to its limit and discards excess assignments without reassignment. Let $\mathcal{A}_p^{\text{patch}}$ denote the retained expert indices for patch p. Router regularization is described in Section 3.4.

3.3 PATCH FEATURE REPLACEMENT

The retained experts construct the feature supplied to token projection. For a nonempty retained set, their outputs are uniformly averaged:

$$\mathcal{R}_{\text{patch}}(\mathbf{h}_p^{\text{patch}}) = \frac{1}{|\mathcal{A}_p^{\text{patch}}|} \sum_{e \in \mathcal{A}_p^{\text{patch}}} \phi_e(\mathbf{h}_p^{\text{patch}}). \tag{1}$$

If no assignment remains, $\mathcal{R}_{\text{patch}}$ returns zero. Token projection receives

$$\overline{\mathbf{u}}_p = \begin{cases} \mathcal{R}_{\text{patch}}(\mathbf{h}_p^{\text{patch}}), & \mathcal{A}_p^{\text{patch}} \neq \varnothing, \\ \mathbf{u}_p, & \mathcal{A}_p^{\text{patch}} = \varnothing, \end{cases} \qquad \mathbf{t}_p = g_{\text{proj}}(\overline{\mathbf{u}}_p). \tag{2}$$

The mean of the retained expert outputs replaces the convolutional feature; the original feature passes through only if all selected assignments are discarded. With residual addition, token projection instead receives $\mathbf{u}_p + \mathcal{R}_{\text{patch}}(\mathbf{h}_p^{\text{patch}})$, so the expert branch supplies an additive correction to the original feature. Section 5.2 compares these integration choices with matched expert architectures, routing rules, and normalization.

The patch feature extractor and token projection remain trainable. A straight-through estimator propagates gradients through retained assignment probabilities while preserving the forward mean. Appendix A.3 specifies selection, capacity filtering, and the estimator.

3.4 ROUTING REGULARIZATION AND TRAINING OBJECTIVE

The routing loss $\mathcal{L}_{\text{route}}$ combines a confidence–diversity penalty (Mustafa et al., 2022) with load-balancing and router-logit penalties. For noise-free routing probabilities $\boldsymbol{\pi}_i$ over $P > 1$ experts, let n denote the number of patch inputs in the routing batch and $\overline{\boldsymbol{\pi}} = n^{-1} \sum_{i=1}^{n} \boldsymbol{\pi}_i$ their mean. With $\mathcal{H}_P$ denoting entropy normalized by $\log P$, we define the confidence–diversity penalty:

$$\mathcal{J}_{\text{patch}}(t) = \beta(t) \frac{1}{n} \sum_{i=1}^{n} \mathcal{H}_P(\boldsymbol{\pi}_i) - \mathcal{H}_P(\overline{\boldsymbol{\pi}}). \tag{3}$$

The first term encourages confident per-patch choices, while the second encourages diverse expert utilization across the batch. Here t indexes training epochs and $\beta(t)$ controls the relative weight of confidence. Statistics are computed before capacity filtering, so inputs with discarded assignments still contribute. Appendix A.4 specifies the weighting and aggregation across encoder passes.

We combine this routing loss with masked-patch reconstruction and contrastive learning to train PLARE jointly:

$$\mathcal{L}_{\text{SSL}} = \mathcal{L}_{\text{gen}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{route}}. \tag{4}$$

Here $\mathcal{L}_{\text{gen}}$ trains a decoder to reconstruct masked waveform patches from contextualized visible-patch representations. $\mathcal{L}_{\text{con}}$ aligns projected CLS representations of unmasked same-patient recordings and their augmented views. The masked reconstruction and unmasked contrastive passes share the encoder; the decoder and contrastive projector are used only during pretraining. Appendices A.5 and B provide the objective definitions and training settings.

4 EXPERIMENTAL SETUP

4.1 DATA AND EVALUATION PROTOCOLS

We pretrained on MIMIC-IV-ECG (Gow et al., 2023), CODE-15% (Ribeiro et al., 2021), and IKEM (Seják et al., 2023). Downstream evaluation used four PTB-XL label sets (Super, Sub, Rhythm, and

Form) (Wagner et al., 2020) and CPSC2018 (Liu et al., 2018). PTB-XL used folds 1–8/9/10 for training/validation/test; CPSC2018 used a 7:1:2 split. Linear probing (LP) trained a linear head on the frozen encoder; fine-tuning (FT) updated both encoder and head. Data and protocol details appear in Appendices B.1 and B.2.

4.2 Baselines and Controlled Comparisons

We pretrained the ten baselines in Tables 1 and 2 on the same unlabeled pool, using common downstream procedures where compatible with their architectures. Superscripts denote ECGFounder–CommonSSL[a], ECGFM–CPC[b], and ECG-FM–WCR[c]; Appendix B.3 details the implementations. Dagger-marked WCR scores for CPSC2018 were computed without constant-lead recordings; these scores are excluded from rankings and Δ references requiring identical test ECGs.

PLARE-Residual uses the same expert architectures, expert count, and router as PLARE but adds the aggregated expert output to the original patch feature with unit weight. This control shares the temporal Transformer, token projection, input processing, and reconstruction and contrastive objectives with PLARE. The control without content-dependent routing (w/o CDR) uses an approximately balanced random mapping from the 82 patch positions in each ten-second input to four experts. The same mapping is used for every ECG throughout training and evaluation, independently of waveform content; specialization by temporal position remains possible (Appendix D.1).

4.3 Implementation and Reporting

The primary configuration used $P = 4$ patch experts with $k_P = 1$ and 12-lead, 10-second inputs at 250 Hz across datasets and protocols. Patches contained 60 time points (240 ms) with a stride of 30 time points (120 ms). Architecture and training details appear in Appendices A and B.5.

AUROC and AUPRC are macro-averaged across labels on a 0–100 scale. They measure discrimination and precision–recall performance, respectively (Saito & Rehmsmeier, 2015). We define $C = (\mathrm{AUROC} + \mathrm{AUPRC})/2$. Section 5.1 reports AUROC/AUPRC in that order; subsequent performance tables use C to conserve space. Detailed component results appear in Appendices C–F.

Where used, boldface marks the best available scores and underlining the second-best; comparisons with two available models highlight only the best, and ties share ranks. In the ablation study, Δ is the ablated variant minus PLARE; elsewhere, it is PLARE minus the best eligible comparator. Differences are computed from displayed scores in percentage points (blue: positive; red: negative).

We selected pretrained PLARE checkpoints with the lowest total validation SSL loss within 200 epochs and downstream checkpoints with the highest validation C. External comparisons used one common downstream seed; controlled comparisons used five downstream seeds per fixed pretrained checkpoint and report mean $\pm$ sample SD of seed-wise C, quantifying downstream variability conditional on that checkpoint. Each Top-k setting was pretrained separately (Appendix D.2).

5 Results

5.1 Comparison with Existing Methods

We first evaluated downstream ECG classification performance. Tables 1 and 2 compare the fixed primary configuration with the external baselines under LP and FT.

Linear probing. PLARE achieved higher AUROC and AUPRC than the baseline methods on Sub, Rhythm, Form, and CPSC. On Super, CREMA achieved the highest AUROC and CREMA-Max the highest AUPRC. Because LP left the encoder frozen, these improvements indicate that the pretrained features supported downstream discrimination without supervised encoder adaptation. The Super results also show that the advantage differed across label sets within the same ECG corpus.

Fine-tuning. PLARE achieved higher AUROC and AUPRC than the baseline methods on Sub, Rhythm, Form, and CPSC, while CREMA achieved the highest values for both metrics on Super. The gains therefore extended to supervised adaptation of the encoder (Appendix C).

Across these four tasks, PLARE achieved larger absolute gains in AUPRC than in AUROC over the strongest eligible baseline for each metric under both protocols. This pattern highlights improved precision–recall performance for positive-case retrieval across both frozen-feature evaluation and supervised encoder adaptation.

Table 1: Comparison with existing methods under LP.

Model	Super	Sub	Rhythm	Form	CPSC
ECGFounder[a]	86.73 / 69.78	89.22 / 38.50	89.54 / 46.54	73.07 / 19.80	89.49 / 62.52
SimCLR	89.24 / 75.97	87.13 / 36.85	83.07 / 36.15	74.75 / 20.49	87.03 / 56.59
CLoCS	89.72 / 76.87	90.43 / 41.48	86.34 / 36.23	77.49 / 24.01	90.24 / 63.32
COMET	90.44 / 78.05	90.57 / 43.82	86.00 / 36.62	75.01 / 22.14	90.28 / 62.95
ECGFM[b]	86.06 / 67.54	82.55 / 29.33	81.78 / 28.46	63.43 / 15.64	79.25 / 38.17
ECG-FM[c]	85.73 / 68.68	89.35 / 39.65	85.30 / 48.39	76.88 / 23.26	89.62[†] / 64.62[†]
ST-MEM	81.49 / 59.90	78.52 / 24.74	82.53 / 34.31	66.25 / 14.04	79.05 / 37.41
MAE	87.55 / 73.45	88.10 / 35.39	87.07 / 43.63	77.39 / 20.77	87.72 / 57.05
CREMA	**91.25** / 78.82	91.37 / 45.99	88.92 / 42.16	80.14 / 22.77	92.78 / 69.61
CREMA-Max	91.05 / **79.30**	92.06 / 47.32	87.18 / 42.81	81.58 / 24.17	92.16 / 69.14
PLARE	91.04 (Δ −0.21)	**92.45** (Δ +0.39)	**93.47** (Δ +3.93)	**82.27** (Δ +0.69)	**94.22** (Δ +1.44)
	78.94 (Δ −0.36)	**51.53** (Δ +4.21)	**52.82** (Δ +4.43)	**26.19** (Δ +2.02)	**75.30** (Δ +5.69)

Table 2: Comparison with existing methods under FT.

Model	Super	Sub	Rhythm	Form	CPSC
ECGFounder[a]	89.64 / 75.61	91.75 / 46.20	95.59 / 50.01	79.91 / 25.28	94.10 / 75.12
SimCLR	92.28 / 81.87	92.04 / 48.40	91.04 / 46.25	84.66 / 32.36	94.33 / 75.35
CLoCS	92.09 / 81.45	90.78 / 50.19	91.14 / 46.57	80.43 / 28.81	93.87 / 75.52
COMET	92.30 / 81.56	92.20 / 50.96	91.88 / 50.33	78.98 / 28.46	94.21 / 74.88
ECGFM[b]	90.63 / 78.10	90.78 / 43.29	94.45 / 53.64	80.59 / 29.91	95.25 / 80.99
ECG-FM[c]	90.76 / 78.59	91.92 / 46.30	95.47 / 54.31	86.89 / 34.57	96.49[†] / 82.65[†]
ST-MEM	87.95 / 72.80	87.84 / 37.74	90.18 / 46.07	72.98 / 22.36	93.20 / 71.18
MAE	92.23 / 81.31	91.12 / 47.05	91.29 / 53.36	85.81 / 32.86	95.04 / 78.69
CREMA	**92.86** / **82.56**	93.35 / 50.20	91.94 / 52.81	87.07 / 32.49	95.67 / 81.15
CREMA-Max	91.91 / 80.87	92.43 / 48.33	90.56 / 48.87	83.70 / 29.54	95.05 / 77.68
PLARE	92.66 (Δ −0.20)	**93.92** (Δ +0.57)	**97.59** (Δ +2.00)	**87.30** (Δ +0.23)	**97.46** (Δ +1.79)
	82.01 (Δ −0.55)	**54.15** (Δ +3.19)	**61.48** (Δ +7.17)	**36.74** (Δ +2.17)	**86.56** (Δ +5.41)

5.2 Ablation Analysis

We first examine the contributions of feature replacement and content-dependent routing, then evaluate the total number of patch experts (P) and the number selected per patch (k_P).

Table 3: Patch feature ablations under LP. Repl.: replacement; CDR: content-dependent routing. The w/o Repl. variant uses residual addition.

Model	Super	Sub	Rhythm	Form	CPSC
PLARE	84.99±0.05	72.21±0.55	71.76±0.83	52.57±1.06	84.70±0.08
w/o Repl.	85.28±0.15 (Δ+0.29)	70.97±0.59 (Δ−1.24)	69.59±0.89 (Δ−2.17)	52.50±0.73 (Δ−0.07)	84.35±0.46 (Δ−0.35)
w/o CDR	83.75±0.01 (Δ−1.24)	71.86±0.50 (Δ−0.35)	69.80±0.81 (Δ−1.96)	51.93±0.54 (Δ−0.64)	83.68±0.18 (Δ−1.02)

Table 4: Patch feature ablations under FT. Repl.: replacement; CDR: content-dependent routing. The w/o Repl. variant uses residual addition.

Model	Super	Sub	Rhythm	Form	CPSC
PLARE	87.31±0.27	73.05±0.86	79.14±0.66	61.32±0.94	91.72±0.34
w/o Repl.	87.05±0.36 (Δ−0.26)	72.08±0.57 (Δ−0.97)	78.81±0.23 (Δ−0.33)	61.25±0.47 (Δ−0.07)	91.83±0.24 (Δ+0.11)
w/o CDR	86.80±0.22 (Δ−0.51)	72.04±0.59 (Δ−1.01)	76.56±0.97 (Δ−2.58)	60.99±0.85 (Δ−0.33)	91.07±0.26 (Δ−0.65)

Replacement versus residual integration. Tables 3 and 4 report PLARE and PLARE-Residual (w/o Repl.), as defined in Section 4.2. Both models used four patch experts with Top-1 selection. Using residual addition instead of feature replacement reduced C on four of five tasks under each protocol. The reductions ranged from 0.07 to 2.17 points under LP and from 0.07 to 0.97 points under FT. Residual addition improved C only on Super under LP (+0.29) and CPSC under FT (+0.11).

With the expert and other component architectures unchanged, these results support replacement as an effective feature-integration component for downstream performance under both protocols.

Effect of CDR (content-dependent routing). Using fixed random assignment instead of content-dependent routing reduced C on all five tasks under each protocol. The reductions ranged from 0.35 to 1.96 points under LP and from 0.33 to 2.58 points under FT. Fixed random assignment reduced C most on Rhythm under LP (-1.96) and FT (-2.58). The LP results indicate that content-dependent selection benefits downstream discrimination with a frozen encoder; the consistent advantage under FT shows that this benefit also extends to supervised encoder adaptation. Together with the integration ablation, these results support both content-dependent expert selection and feature replacement as effective components of PLARE.

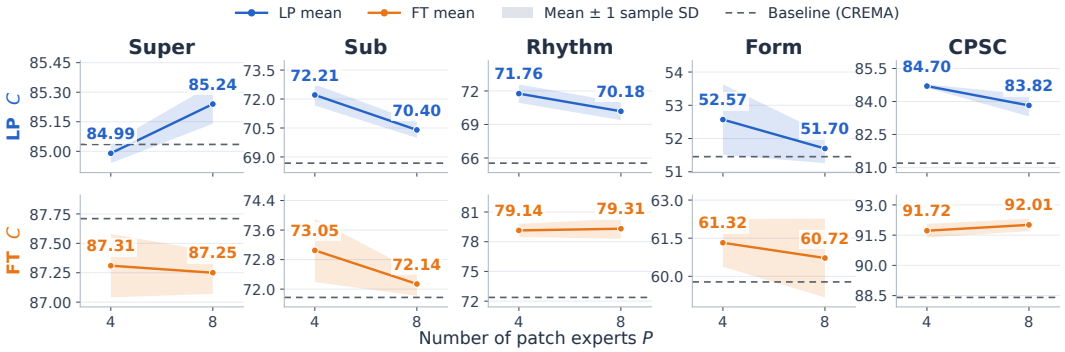

Figure 3: Effect of the total patch expert count ($k_P = 1$). Curves show mean C; shading shows ± 1 sample SD over five downstream seeds per fixed pretrained checkpoint. Dashed lines show CREMA.

Number of patch experts (P). With Top-1 selection ($k_P = 1$), $P = 4$ achieved higher mean C than $P = 8$ on four of five tasks under LP and three under FT (Figure 3). Thus, with one expert selected per patch, increasing the number of experts alone did not consistently improve downstream performance.

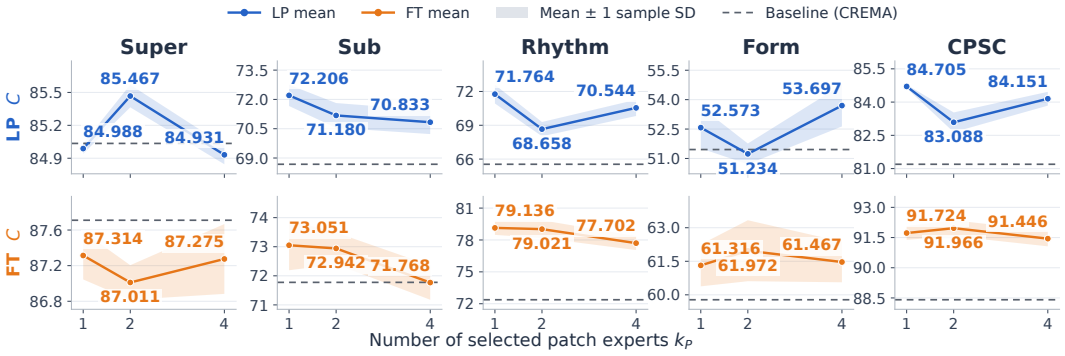

Figure 4: Effect of the selected patch expert count ($P = 4$). Curves show mean C; shading shows ± 1 sample SD over five downstream seeds per fixed pretrained checkpoint. Dashed lines show CREMA.

Number of selected patch experts (k_P). As shown in Figure 4, Top-1 selection achieved the highest C on three of five tasks under each protocol. It exceeded all-expert aggregation ($k_P = 4$) on four of five tasks under both LP and FT (Appendix D.2). Higher selection counts performed best on Super and Form under LP and on Form and CPSC under FT. Thus, increasing the number of experts selected per patch did not consistently improve downstream performance; the selection count yielding the highest mean C varied across tasks and protocols.

5.3 PERFORMANCE AND INFERENCE COST

We measured inference cost to place the LP performance gains in a computational context. Table 5 reports LP C averaged over the four PTB-XL tasks with parameter count, FLOPs, latency, and peak

memory. Profiling used an A100-SXM4-40GB GPU, FP32, and batch size one at each model's input configuration. Each of ten processes measured median latency over 600 inferences. Active FLOPs count shared computation and retained expert transformations; total FLOPs include all expert transformations as a computational reference. Appendix F gives the full procedure.

Table 5: Inference costs and LP performance (averaged over the four PTB-XL tasks in Table 1). Latency and memory are means over ten processes. PLARE uses $P = 4$, $k_P = 1$.

Model	Input (Hz, s)	Params (M)	GFLOPs Active / Total	Latency (ms)	Peak memory (GB)	LP C
CREMA	250, 10	18.041	3.518 / 3.518	2.190	0.0883	67.680
CREMA-Max	250, 10	177.601	30.586 / 30.586	6.917	0.7484	<u>68.185</u>
ECG-FM–WCR	500, 5	90.887	29.303 / 29.303	5.452	0.4106	64.660
PLARE	250, 10	69.528	8.686 / 12.318	4.611	0.2959	**71.090** (Δ+2.905)

PLARE achieved higher LP C than CREMA-Max and ECG-FM–WCR with fewer parameters, lower FLOPs, lower peak memory, and lower latency at their respective inputs. Relative to CREMA-Max, it used 39% of the parameters, and its latency was 67% of that of CREMA-Max. CREMA was smaller and faster, with lower C. These comparisons characterize PLARE's measured performance–cost tradeoff.

5.4 EXPERT ALLOCATION ANALYSIS

We examined whether the pretrained router selects different experts across patches and whether its selections are associated with normal (N) and premature ventricular contraction (PVC; V) beat contexts. Using INCART's beat annotations (Yakushenko, 2008), we combined statistical analyses with individual-input examples to characterize overall expert selection frequencies and their variation with local waveform context.

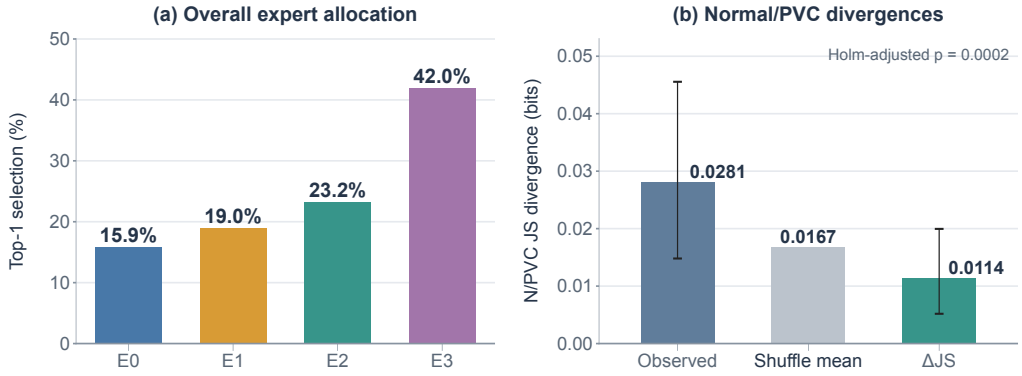

Figure 5: INCART expert allocation: (a) overall selection; (b) Normal/PVC divergences.

Overall allocation and normal/PVC differences. Using the primary configuration (Section 4.3) and the frozen pretrained checkpoint used downstream (epoch 75), we recorded noise-free router probabilities and Top-1 selections before capacity filtering over 13,500 windows from 75 recordings. For each beat, we averaged selection indicators over patches centered in the interval bounded by midpoints to neighboring annotations and padded by 120 ms per side, then averaged these vectors separately for N and V within each recording. Across 68 recordings from 30 patients, we compared patient-averaged Jensen–Shannon (JS) divergence with a null distribution from 10,000 within-window label permutations preserving N/V counts, waveforms, and routing; error bars show 95% patient-bootstrap intervals (Appendix G).

E0–E3 had overall selection fractions of 15.9%, 19.0%, 23.2%, and 42.0%, respectively (Figure 5a). Observed JS divergence was 0.0281 bits versus a permutation mean of 0.0167 bits (Figure 5b), giving an excess $\Delta\text{JS} = 0.0114$ bits (95% CI $[0.0052, 0.0200]$; Holm-adjusted $p = 0.0002$). Thus, the router selected all four experts at different frequencies, while N/PVC allocation differed beyond the within-window reference. These results link local expert selection to annotated beat context in the frozen pretrained encoder.

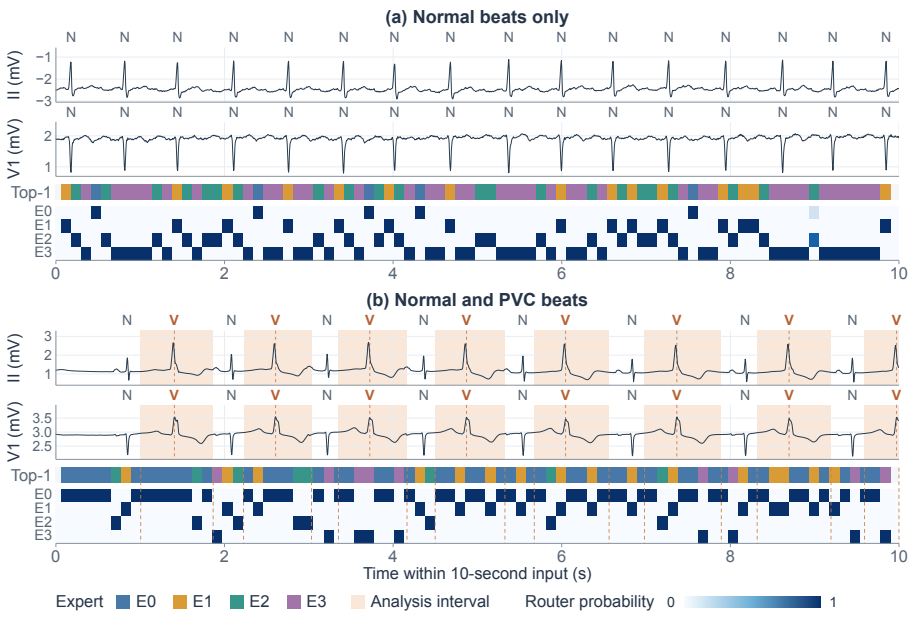

Figure 6: Expert allocation in (a) I50, 1550–1560 s (16 N, no V), and (b) I32, 480–490 s (8 N, 8 V). Beneath leads II and V1, Top-1 strips and probability heatmaps display routing at patch centers. Shading marks padded V-beat intervals, clipped only for display; dashed lines mark V annotations in waveforms and interval boundaries in routing panels.

Allocation within individual inputs. Using the same routing outputs, we inspected two windows selected post hoc. Each input yielded 82 patches (240 ms width; 120 ms stride). E3 (pink) was selected most frequently in the normal-only window (Figure 6a), whereas E0 (blue) was selected most often in the mixed normal/PVC window and repeatedly near V annotations (Figure 6b). These two examples suggest possible differences in expert selection between normal-beat and PVC-related contexts. In the recordings retained for the Normal/PVC analysis, normal beats substantially outnumbered PVCs (Appendix G). Given this imbalance, the higher overall selection frequency of E3 than E0 (Figure 5a) is consistent with the selection patterns in these two examples. Together with the Normal/PVC allocation differences (Figure 5b), these observations suggest context-dependent expert allocation across local waveform contexts; however, these results do not establish fixed clinical roles for individual experts.

5.5 Contextual Expert Extension

We investigated whether contextual experts provide further gains beyond patch-level processing. PLARE-Joint retains the patch-level MoE and replaces standard FFNs in later temporal Transformer blocks with four-expert MoEs, using Top-1 selection throughout. PLARE achieved higher mean C on three of five tasks under LP and four under FT (Appendix E, Table 16), showing no consistent benefit from this extension under the evaluated setting.

6 Conclusion

We introduced PLARE, which constructs local ECG patch features through content-dependent routing and feature replacement before temporal contextualization. It achieved higher AUROC and AUPRC than the evaluated baselines on four of five tasks under both LP and FT. In ablations under both protocols, content-dependent routing achieved higher mean C than fixed random assignment on all five tasks, while replacement achieved higher mean C than residual addition on four. With Top-1 patch routing, PLARE showed a favorable measured performance–cost tradeoff against larger baselines at their respective input configurations. On INCART, the frozen router selected all four experts, with selection distributions associated with annotated beat context. Together, these gains, controlled ablations, and allocation patterns support using content-dependent routing and feature replacement to construct local patch features for self-supervised ECG representation learning.

AI USE STATEMENT

Generative AI tools were used to aid or polish the writing of this manuscript. The authors take full responsibility for the final text, methods, results, and claims.

REFERENCES

Sana M. Al-Khatib, William G. Stevenson, Michael J. Ackerman, William J. Bryant, David J. Callans, et al. 2017 AHA/ACC/HRS guideline for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: A report of the american college of cardiology/american heart association task force on clinical practice guidelines and the heart rhythm society. *Circulation*, 138(13):e272–e391, 2018. doi: 10.1161/CIR.0000000000000549.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation, 2013. URL https://arxiv.org/abs/1308.3432.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, 2020.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 2022.

Kun Feng, Shaocheng Lan, Yuchen Fang, Wenchao He, Sihan Lu, Shuqi Gu, Lintao Ma, Xingyu Lu, and Kan Ren. Kairos: Toward adaptive and parameter-efficient time series foundation models, 2025. URL https://arxiv.org/abs/2509.25826. Version 3, revised May 2026.

C. A. Field and A. H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):369–390, 2007. doi: 10.1111/j.1467-9868.2007.00593.x.

Brian Gow, Tom Pollard, Larry A. Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W. Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic electrocardiogram matched subset. PhysioNet, 2023. URL https://physionet.org/content/mimic-iv-ecg/1.0/. Version 1.0.

İnan Güler and Elif Derya Übeyli. A modified mixture of experts network structure for ECG beats classification with diverse features. *Engineering Applications of Artificial Intelligence*, 18(7): 845–856, 2005. doi: 10.1016/j.engappai.2005.03.002.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Yu Hen Hu, Sandhya Palreddy, and Willis J. Tompkins. A patient-adaptable ECG beat classifier using a mixture of experts approach. *IEEE Transactions on Biomedical Engineering*, 44(9):891–900, 1997. doi: 10.1109/10.623058.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. Deep networks with stochastic depth, 2016. URL https://arxiv.org/abs/1603.09382.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

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

Fred M. Kusumoto, Mark H. Schoenfeld, Coletta Barrett, James R. Edgerton, Kenneth A. Ellenbogen, et al. 2018 ACC/AHA/HRS guideline on the evaluation and management of patients with bradycardia and cardiac conduction delay: A report of the american college of cardiology/american heart association task force on clinical practice guidelines and the heart rhythm society. *Circulation*, 140 (8):e382–e482, 2019. doi: 10.1161/CIR.0000000000000628.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. MIRA: Medical time series foundation model for real-world health data. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-3296.

Jun Li, Aaron D. Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M. Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *NEJM AI*, 2(7):AIoa2401033, 2025b. doi: 10.1056/AIoa2401033.

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*, 2018.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL `https://arxiv.org/abs/1608.03983`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://arxiv.org/abs/1711.05101`.

Nastaran Mansourian, Arash Mohammadi, M. Omair Ahmad, and M. N. S. Swamy. ECG-emotionnet: Nested mixture of expert (NMoE) adaptation of ECG-foundation model for driver emotion recognition. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–6, 2025. doi: 10.1109/EMBC58623.2025.11254389.

Kaden McKeen, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. ECG-FM: An open electrocardiogram foundation model. *JAMIA Open*, 8(5):ooaf122, 2025. doi: 10.1093/jamiaopen/ooaf122.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with LIMoE: the language-image mixture of experts. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0695.

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

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1:792–795, 2026. doi: 10.1038/s44360-026-00096-z.

Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023.

Antônio H. Ribeiro, Gabriela M.M. Paixao, Emilly M. Lima, Manoel Horta Ribeiro, Marcelo M. Pinto Filho, Paulo R. Gomes, Derick M. Oliveira, Wagner Meira Jr, Thömas B Schon, and Antonio Luiz P. Ribeiro. Code-15%: A large scale annotated dataset of 12-lead ecgs. Zenodo, 2021.

Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Michal Seják, Jakub Sido, and David Žahour. Ikem dataset v1.0.0. Zenodo, 2023.

Daniel Shao, Joel Runevic, Richard Chen, Drew Williamson, Ahrong Kim, Andrew Song, and Faisal Mahmood. Mixture of mini experts: Overcoming the linear layer bottleneck in multiple instance learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/87425754bcc35f2bc62ef4a421a772d6-Abstract-Conference.html.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *Proceedings of the International Conference on Learning Representations*, 2025.

Junho Song, Jong-Hwan Jang, DongGyun Hong, Joon-myoung Kwon, and Yong-Yeon Jo. CREMA: A contrastive regularized masked autoencoder for robust ECG diagnostics across clinical domains, 2024. Version 3, revised 21 August 2025.

Diogo Tuler, Pedro Robles Dutenhefner, Jose Geraldo Fernandes, Turi Rezende, Gabriel Lemos, Gisele L. Pappa, Gabriela Paixão, Antônio Ribeiro, and Wagner Meira Jr. Leveraging cardiologists prior-knowledge and a mixture of experts model for hierarchically predicting ECG disorders. In *Computing in Cardiology*, volume 51, 2024. doi: 10.22489/CinC.2024.477.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018. URL https://arxiv.org/abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

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*, 2020.

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

Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. Contrast everything: A hierarchical contrastive framework for medical time-series. *Advances in Neural Information Processing Systems*, 2023.

Evgeny Yakushenko. St petersburg INCART 12-lead arrhythmia database. PhysioNet, version 1.0.0, 2008. URL https://physionet.org/content/incartdb/1.0.0/.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training BERT in 76 minutes. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.00962.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html.

Guodao Zhang, Cunnan Wei, Yanjie Lu, and Hong Sun. MoViE: A mixture-of-experts multimodal fusion model for cardiovascular disease detection. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 782–787, 2025a. doi: 10.1109/SMC58881.2025.11342849.

Shaoting Zhang, Yishan Du, Wenji Wang, Xianying He, Fangfang Cui, Liang Zhao, Bei Wang, Zhiqiang Hu, Ziqiang Wang, Qing Xia, et al. Ecgfm: A foundation model for ecg analysis trained on a multi-center million-ecg dataset. *Information Fusion*, 2025b.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 2022.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models, 2022. URL https://arxiv.org/abs/2202.08906.

A Model Architecture and Learning Objectives

This appendix specifies the primary PLARE architecture, routing operations, and learning objectives. The primary model uses four patch experts with Top-1 selection and standard FFNs in all six Transformer blocks. The contextual extension is described in Appendix A.6.

Here P is the number of patch experts and S is the number in each contextual module; zero denotes the corresponding shared transformation. The selection counts are k_P and k_S. MoE modules are indexed by $\mathcal{M} \subseteq \{\text{patch}, B_{\lfloor N/2 \rfloor+1}, \ldots, B_N\}$, where B_ℓ denotes Transformer block ℓ. Each module has its own router and experts. For module s, E_s is its expert count and d_s its input width. Thus $(E_{\text{patch}}, d_{\text{patch}}) = (P, F)$, whereas $(E_{B_\ell}, d_{B_\ell}) = (S, D)$ for a contextual expert module.

A.1 Architectural Specification

Table 6: Primary PLARE architecture ($P = 4$, $k_P = 1$, $N = 6$).

Component	Specification
Input	12 leads, 10 s, 250 Hz
Patch feature extractor	Convolutional stem and residual blocks (Conv/Res)
Patch-level MoE	Content-dependent routing and patch feature replacement
Token projection	Patch features ($F = 1920$) to tokens ($D = 1024$)
Temporal Transformer	6 blocks with standard FFNs
Routing rule	Noisy Top-k; uniform aggregation with straight-through gradients
Pretraining objectives	Masked reconstruction + unmasked multi-relation contrastive
Downstream representation	Final CLS representation from the unmasked input

Dimensions and expert networks. Patches contain 60 samples (240 ms) with stride 30 (120 ms), yielding 82 patches per 10-second record. The patch feature extractor comprises a 32-channel convolutional stem with kernel width 15 and two residual stages with kernel width 3 and two residual blocks per stage. The flattened patch features have width $F = 1920$ and are projected to $D = 1024$; each Transformer block has 16 attention heads. Patch experts are $1920 \to 1920 \to 1920$ MLPs with ReLU. The reconstruction decoder has two blocks, width 512, and 16 heads. The contrastive projector has two hidden layers and an output projection, all of width 1024, followed by unit-length normalization.

A.2 ECG Encoder Formulation

Let $\mathbf{x}_b \in \mathbb{R}^{L \times T}$ denote ECG record b, with L leads and T samples per lead, in a minibatch of size B. A window of width w and stride δ produces $M = 1 + \lfloor (T - w)/\delta \rfloor$ patches $\mathbf{x}_{b,p} \in \mathbb{R}^{L \times w}$. The shared patch feature extractor, including flattening, produces

$$\mathbf{u}_{b,p} = f_{\text{patch}}(\mathbf{x}_{b,p}) \in \mathbb{R}^F. \tag{5}$$

Let LN denote layer normalization (Ba et al., 2016). The patch-level MoE input is $\mathbf{h}_{b,p}^{\text{patch}} = \text{LN}_{\text{patch,in}}(\mathbf{u}_{b,p})$. After the replacement in Equation 2, token projection uses layer normalization and root-mean-square normalization (Zhang & Sennrich, 2019) (RMSNorm):

$$\mathbf{t}_{b,p} = \text{RMSNorm}_{\text{patch,proj}}(\mathbf{W}_{\text{proj}}\, \text{LN}_{\text{patch,out}}(\overline{\mathbf{u}}_{b,p}) + \mathbf{b}_{\text{proj}}) \in \mathbb{R}^D. \tag{6}$$

For masking ratio ρ, let $\boldsymbol{\sigma}_b = (\sigma_{b,1}, \ldots, \sigma_{b,M_\rho})$ be the ordered visible-patch indices, where $M_\rho = \lfloor M(1 - \rho) \rfloor$. With fixed positional embedding $\mathbf{e}_p$ and learned CLS vector $\mathbf{c}_{\text{cls}}$, the temporal Transformer input is

$$\mathbf{Z}_0^{(b)} = \left[\mathbf{c}_{\text{cls}}; \left(\mathbf{t}_{b,\sigma_{b,j}} + \mathbf{e}_{\sigma_{b,j}}\right)_{j=1}^{M_\rho}\right]. \tag{7}$$

The masking ratio ρ is positive only for masked reconstruction and is zero for contrastive learning and downstream evaluation. For masked reconstruction, patch routing precedes random masking: all patches contribute to capacity and auxiliary statistics, whereas only visible patch tokens enter the temporal Transformer. The Transformer (Vaswani et al., 2017) applies

$$\mathbf{A}_\ell = \mathbf{Z}_{\ell-1} + \text{DropPath}_{\ell,\text{attn}}(\text{MSA}_\ell(\text{RMSNorm}_{\ell,1}(\mathbf{Z}_{\ell-1}))), \tag{8}$$

$$\mathbf{Z}_\ell = \mathbf{A}_\ell + \text{DropPath}_{\ell,\text{ffn}}(\Psi_\ell(\text{RMSNorm}_{\ell,2}(\mathbf{A}_\ell))), \qquad \ell = 1, \ldots, N. \tag{9}$$

Here, $\mathrm{DropPath}$ denotes stochastic depth (Huang et al., 2016). In PLARE-Joint, Ψ_ℓ denotes the corresponding independent contextual MoE module for $\ell > \lfloor N/2 \rfloor$; PLARE uses a standard FFN in every block. Final RMS normalization of $\mathbf{Z}_N$ yields $\widetilde{\mathbf{Z}}^{(b)} = [\mathbf{z}_{\mathrm{cls}}^{(b)}; \mathbf{Z}_{\mathrm{patch}}^{(b)}]$, with patch tokens ordered by $\boldsymbol{\sigma}_b$.

A.3 Expert Routing and Gradient Propagation

Each MoE module selects experts for an input and averages their outputs after capacity filtering. Consider module $s \in \mathcal{M}$ with input $\mathbf{h}_i^s \in \mathbb{R}^{d_s}$, bias-free router $\mathbf{W}_r^s \in \mathbb{R}^{E_s \times d_s}$, and expert transformations $\phi_e^s : \mathbb{R}^{d_s} \to \mathbb{R}^{d_s}$. The noise-free router logits are $\mathbf{g}_i^s = \mathbf{W}_r^s \mathbf{h}_i^s$. The patch-level MoE uses $(E_s, k_s) = (P, k_P)$, whereas each contextual MoE module uses $(E_s, k_s) = (S, k_S)$, with $1 \le k_s \le E_s$. At routing temperature $\vartheta(t) > 0$, let $\boldsymbol{\pi}_i^s$ denote the noise-free probabilities and $\widetilde{\boldsymbol{\pi}}_i^s$ the probabilities used for expert selection:

$$\boldsymbol{\pi}_i^s = \mathrm{softmax}\left(\frac{\mathbf{g}_i^s}{\vartheta(t)}\right), \qquad \widetilde{\boldsymbol{\pi}}_i^s = \mathrm{softmax}\left(\frac{\mathbf{g}_i^s + \boldsymbol{\epsilon}_i^s}{\vartheta(t)}\right), \qquad \mathcal{T}_i^s = \mathrm{TopK}(\widetilde{\boldsymbol{\pi}}_i^s, k_s). \tag{10}$$

During training, noisy Top-k routing (Shazeer et al., 2017) uses $\boldsymbol{\epsilon}_i^s = \gamma(t)\boldsymbol{\nu}_i^s \odot \boldsymbol{\xi}_i^s$, where $\boldsymbol{\xi}_i^s \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{E_s})$ and $\boldsymbol{\nu}_i^s = \mathrm{softplus}(\mathbf{W}_n^s \mathbf{h}_i^s + \mathbf{b}_n^s) + \varepsilon_{\mathrm{noise}}$. The positive constant $\varepsilon_{\mathrm{noise}}$ prevents a zero noise scale. The noise parameters $\mathbf{W}_n^s, \mathbf{b}_n^s$ are learned; $\gamma(t)$ scales the noise, and $\boldsymbol{\epsilon}_i^s = \mathbf{0}$ at inference.

For n_s module inputs, the capacity of each expert is $\mathrm{cap}_s = \lceil n_s k_s \alpha_{\mathrm{cap}} / E_s \rceil$ with $\alpha_{\mathrm{cap}} \ge 1$. Each expert retains its highest-scoring assignments up to this limit. The retained experts for input i form the set $\mathcal{A}_i^s \subseteq \mathcal{T}_i^s$; excess assignments are discarded without reassignment. The forward output is the mean of the retained expert outputs, or zero if no assignments remain:

$$\mathcal{R}_s(\mathbf{h}_i^s) = \begin{cases} \dfrac{1}{|\mathcal{A}_i^s|} \displaystyle\sum_{e \in \mathcal{A}_i^s} \phi_e^s(\mathbf{h}_i^s), & \mathcal{A}_i^s \neq \varnothing, \\ \mathbf{0}, & \mathcal{A}_i^s = \varnothing. \end{cases} \tag{11}$$

Here $n_{\mathrm{patch}} = B'M$ and $n_{B_\ell} = B'(1 + M_\rho)$; $B' = B$ for reconstruction and $B' = 2B$, $\rho = 0$ for contrastive learning. This expert capacity is computed separately per MoE module and objective.

For $m_i^s = |\mathcal{A}_i^s| > 0$ and $a_{i,e}^s = \mathbf{1}\{e \in \mathcal{A}_i^s\}$, we use a straight-through surrogate (Bengio et al., 2013) with coefficient $q_{i,e}^s = a_{i,e}^s / m_i^s + a_{i,e}^s(\widetilde{\pi}_{i,e}^s - \mathrm{sg}(\widetilde{\pi}_{i,e}^s))$, where sg denotes stop-gradient. The sum $\sum_e q_{i,e}^s \phi_e^s(\mathbf{h}_i^s)$ preserves the forward mean in Equation 11 and propagates router gradients through retained assignment probabilities. Only retained expert outputs ($a_{i,e}^s = 1$) receive direct gradients from reconstruction, contrastive learning, or supervised prediction.

At $k_s = E_s$, $\mathrm{cap}_s \ge n_s$, so all assignments are retained: every expert contributes to each input's uniform average and receives these direct gradients. Both Top-1 and all-expert selection use matched expert architectures and counts and retain straight-through router learning. Their regularization losses are defined in Appendix A.4.

A.4 Routing Regularization Losses

The auxiliary routing losses regularize the noise-free logits and probabilities defined in Appendix A.3, before capacity filtering. They use all inputs to each module, including inputs with discarded assignments. Entropy statistics are computed over each module's input batch on each worker, without morphology-frequency estimates or rarity-dependent weights.

For a distribution $\mathbf{q}$ over $E_s > 1$ experts, normalized entropy is $\mathcal{H}_s(\mathbf{q}) = -\sum_e q_e \log q_e / \log E_s$. Numerical evaluation uses $\log(q_e + \varepsilon_{\mathrm{ent}})$ with $\varepsilon_{\mathrm{ent}} > 0$. Let $\boldsymbol{\pi}_i^s$ denote the noise-free routing probabilities over E_s experts for input i in a routing batch of n_s inputs, and let $\overline{\boldsymbol{\pi}}^s = n_s^{-1} \sum_{i=1}^{n_s} \boldsymbol{\pi}_i^s$ be their batch mean. For each enabled module, Equation 3 generalizes to $\mathcal{J}_s(t) = \beta(t) n_s^{-1} \sum_{i=1}^{n_s} \mathcal{H}_s(\boldsymbol{\pi}_i^s) - \mathcal{H}_s(\overline{\boldsymbol{\pi}}^s)$. Combining mean conditional entropy and negative marginal entropy encourages low-entropy routing for individual inputs and diverse expert utilization across the batch, following entropy-regularized MoE learning (Mustafa et al., 2022).

Let $f_e^s = n_s^{-1} \sum_i \mathbf{1}\{e = \arg\max_j \pi_{i,j}^s\}$ be the noise-free Top-1 assignment fraction. The load-balancing loss (Fedus et al., 2022) and router z-loss (Zoph et al., 2022) are

$$\mathcal{B}_s = E_s \sum_{e=1}^{E_s} f_e^s \overline{\pi}_e^s, \qquad \mathcal{Z}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} \left(\log \sum_{e=1}^{E_s} \exp g_{i,e}^s \right)^2. \tag{12}$$

The weighted sum of these three penalties gives the routing loss for module s:

$$\mathcal{Q}_s(t) = a(t)\left[\lambda_{\text{bal}}\mathcal{B}_s + \mathcal{J}_s(t)\right] + \lambda_z \mathcal{Z}_s. \tag{13}$$

The weights λ_{bal} and λ_z control balancing and logit stabilization. For the patch-level MoE, $E_{\text{patch}} = P$ and Equation 3 defines $\mathcal{J}_{\text{patch}}$ using entropy normalized by $\log P$; contextual modules use the corresponding entropy over S experts.

Let $\mathcal{Q}_s^{(v)}$ denote this module loss in the reconstruction or contrastive encoder pass, $v \in \{\text{gen}, \text{con}\}$. For PLARE, the total is $\sum_v \mathcal{Q}_{\text{patch}}^{(v)}$. For PLARE-Joint, it is

$$\mathcal{L}_{\text{route}} = \sum_{v\in\{\text{gen},\text{con}\}} \left(\mathcal{Q}_{\text{patch}}^{(v)} + \frac{1}{\lceil N/2 \rceil} \sum_{\ell=\lfloor N/2 \rfloor + 1}^{N} \mathcal{Q}_{B_\ell}^{(v)} \right). \tag{14}$$

For each pass, we add the patch penalty to the mean over enabled contextual MoE modules; disabled modules contribute zero. We sum the reconstruction and contrastive pass penalties. The three contrastive relations (Appendix A.5) share one pass, so its routing penalty is counted once. Fine-tuning adds the corresponding single-pass penalty to the supervised objective.

A.5 Reconstruction and Contrastive Objectives

Following masked autoencoding (He et al., 2022), reconstruction is evaluated on masked patches. Let $\zeta_{b,p} = 1$ when patch p is masked and zero otherwise, and let $\widehat{\mathbf{x}}_{b,p} \in \mathbb{R}^{d_x}$ be its decoder prediction, where $d_x = Lw$. The reconstruction loss is

$$\mathcal{L}_{\text{gen}} = \frac{\sum_{b=1}^{B}\sum_{p=1}^{M} \zeta_{b,p} \left\|\text{vec}(\mathbf{x}_{b,p}) - \widehat{\mathbf{x}}_{b,p}\right\|_1}{d_x \sum_{b=1}^{B}\sum_{p=1}^{M} \zeta_{b,p} + \varepsilon_0}, \qquad \varepsilon_0 > 0. \tag{15}$$

For normalized embedding batches $\mathbf{Y} = (\mathbf{y}_i)_{i=1}^{J}$ and $\mathbf{Y}' = (\mathbf{y}'_i)_{i=1}^{J}$, we use symmetric InfoNCE (van den Oord et al., 2018) with contrastive temperature $\tau > 0$:

$$\mathcal{L}_{\text{NCE}}(\mathbf{Y}, \mathbf{Y}') = -\frac{1}{2J}\sum_{i=1}^{J}\left[\log \frac{e^{\mathbf{y}_i^\top \mathbf{y}'_i/\tau}}{\sum_{j=1}^{J} e^{\mathbf{y}_i^\top \mathbf{y}'_j/\tau}} + \log \frac{e^{\mathbf{y}_i^\top \mathbf{y}'_i/\tau}}{\sum_{j=1}^{J} e^{\mathbf{y}_j^\top \mathbf{y}'_i/\tau}}\right]. \tag{16}$$

Each batch contains $B = 2J$ records arranged into J same-patient pairs (U, V). Let $\mathbf{Y}_U$, $\mathbf{Y}_V$, $\mathbf{Y}_{\widetilde{U}}$, and $\mathbf{Y}_{\widetilde{V}}$ be the normalized projected CLS embeddings of the paired records and their augmented views. The contrastive term, weighted by λ_{con} in Equation 4, is

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{NCE}}(\mathbf{Y}_U, \mathbf{Y}_V) + \mathcal{L}_{\text{NCE}}(\mathbf{Y}_U, \mathbf{Y}_{\widetilde{U}}) + \mathcal{L}_{\text{NCE}}(\mathbf{Y}_U, \mathbf{Y}_{\widetilde{V}}). \tag{17}$$

A.6 Contextual Expert Extension

PLARE-Joint retains the patch-level MoE and uses independent contextual expert FFNs in blocks 4–6. Each contextual expert is a $1024 \to 1024 \to 1024$ MLP with GELU. The initial Transformer blocks produce contextual representations, and expert FFNs refine them in the latter half. Let $\mathbf{A}_\ell$ denote the post-attention state including the attention residual, and $\mathbf{Z}_\ell$ the block output. Omitting normalization and stochastic depth, specified in Appendix A.2,

$$\mathbf{Z}_\ell = \mathbf{A}_\ell + \Psi_\ell(\mathbf{A}_\ell), \qquad \Psi_\ell = \begin{cases} \text{FFN}_\ell, & 1 \le \ell \le \lfloor N/2 \rfloor, \\ \mathcal{R}_{B_\ell}, & \lfloor N/2 \rfloor < \ell \le N. \end{cases} \tag{18}$$

PLARE uses $\Psi_\ell = \text{FFN}_\ell$ in every block. In PLARE-Joint, $\mathcal{R}_{B_\ell}$ applies expert transformations token-wise to representations of the learned classification (CLS) token and patch tokens after self-attention. Unlike patch feature replacement, these outputs are residual updates that preserve the contextual state. Final normalization produces the CLS representation $\mathbf{z}_{\text{cls}}$ and patch representations $\mathbf{Z}_{\text{patch}}$.

B Training and Evaluation Procedures

This appendix first describes the datasets and common training and evaluation protocols, followed by baseline implementations and PLARE-specific sampling and training settings.

B.1 Datasets and Sample Counts

Data composition. Table 7 summarizes candidate training and SSL validation pools from MIMIC-IV-ECG (Gow et al., 2023), CODE-15% (Ribeiro et al., 2021), and IKEM (Seják et al., 2023), together with paired-sampling counts for a representative PLARE-Joint run. Recordings with missing metadata or NaN values in the waveforms were excluded. Patient IDs are corpus-specific; totals therefore count identifiers rather than unique individuals across corpora. Within each corpus, patients were shuffled using a fixed corpus-specific seed and partitioned 90/5/5 into train/validation/test. All recordings assigned to one patient identifier remained in the same split.

Table 7: Three pretraining corpora: candidate pools before pairing (a) and paired-sampling counts for the representative PLARE-Joint run ($P = S = 4$, Top-1) (b). In (b), training counts are upper bounds and SSL validation counts are exact.

Partition	Count	MIMIC-IV	CODE-15%	IKEM	Total
(a) Candidate pool before pairing					
Train	**ECGs**	710,828	305,755	88,106	1,104,689
	Patient IDs	144,776	209,321	27,261	381,358
Validation	**ECGs**	38,458	17,006	5,148	60,612
	Patient IDs	8,043	11,629	1,515	21,187
(b) Paired sampling: unique ECGs and patient IDs					
Train ($\leq$)	**ECGs**	621,042	139,656	71,718	832,416
	Patient IDs	93,085	58,321	17,157	168,563
Validation	**ECGs**	33,408	7,820	4,282	45,510
	Patient IDs	5,111	3,302	984	9,397

Split integrity. Record, source-row, and corpus-specific patient identifiers were disjoint across splits; record and source-row identifiers were also unique within each split.

PLARE's paired-sampling procedure and repeated-input counts are detailed in Appendix B.4.

Downstream datasets. We evaluated downstream performance on PTB-XL Super, Sub, Rhythm, and Form (Wagner et al., 2020), which share an ECG corpus but use different label sets, and CPSC2018 (Liu et al., 2018). PTB-XL folds 1–8/9/10 defined the train/validation/test partitions; all models used the CPSC2018 7:1:2 split, with the WCR exclusions described below. Table 8 lists CREMA's published PTB-XL reference partition counts and the shared CPSC2018 partitions before WCR exclusions. The exact ECG counts for the label-level and paired analyses are reported with those results in Appendix C.

Table 8: Downstream reference partitions: label counts and train/validation/test ECG counts, following Song et al. (2024). PTB-XL entries reproduce the published reference counts; analysis-specific counts appear in Appendix C. WCR's constant-lead exclusions are described below.

Dataset	Labels	Train	Validation	Test
PTB-XL Super	5	17,111	2,156	2,163
PTB-XL Sub	23	17,111	2,156	2,163
PTB-XL Rhythm	12	16,853	2,109	2,103
PTB-XL Form	19	7,202	904	882
CPSC2018	9	4,520	646	1,291

Data consistency across models. Models shared the downstream split assignments, with the WCR exclusions below. Model-specific input processing is described in Appendix B.2.

ECG-FM–WCR excluded 17 unique ECGs with constant leads: one PTB-XL training ECG shared by the four tasks, and 11/4/1 CPSC train/validation/test ECGs. Its CPSC train/validation/test partitions therefore retained 4,509/642/1,290 ECGs. The original split assignments were retained, with no prediction-based filtering. Its CPSC performance estimates are marked with daggers and excluded from rankings based on the same test ECGs because this quality-control step changes the evaluated ECGs. Paired analyses were restricted to models with matched saved predictions; none are reported for the three foundation-model variants.

B.2 Common Training and Evaluation Protocols

Input processing. Input formats were standardized where compatible with each method's objectives and architecture. Preprocessing followed the original papers where specified and the PLARE protocol otherwise. The shared PLARE downstream protocol used 12-lead, 10-second ECGs sampled at 250 Hz, without additional normalization. Processing proceeded in the following order: missing-lead filling, resampling, length adjustment with constant padding, Butterworth bandpass filtering, and powerline removal at the configured frequencies. No dataset-level augmentation was configured for LP or FT.

The three adapted foundation models retained 12-lead, 500 Hz inputs in both pretraining and downstream evaluation: ECGFounder–CommonSSL and ECGFM–CPC used the full 10 seconds, whereas ECG-FM–WCR used the first 5 seconds after normalizing the full 10-second recording. Other baselines shared PLARE's input processing and LP/FT evaluation procedures.

Within each dataset, evaluation protocol, and pretraining condition, PLARE and PLARE-Joint shared recording and window identifiers, labels and their ordering, and input processing. PTB-XL analyses also used the same software dependencies and available patient mappings. LP–FT comparisons reflect differences in encoder adaptation, prediction heads, training budgets, and input filters (Table 9).

Training and checkpoint selection. Models used their respective objectives and architectures. Table 9 summarizes the shared hyperparameter optimization (HPO) search space, including preprocessing candidates, and epoch budgets. Optimizer candidates were LAMB (You et al., 2020), AdamW (Loshchilov & Hutter, 2019), and Adam (Kingma & Ba, 2015). Pretraining used a maximum of 200 epochs, with checkpoints selected by the lowest validation pretraining loss. For PLARE, this criterion was the total $\mathcal{L}_{\text{SSL}}$. PTB-XL downstream experiments used AdamW and binary cross-entropy. LP and FT independently selected downstream checkpoints by maximum validation C (Section 4.3); test data were used only for evaluation.

Table 9: Shared HPO search space and epoch budgets. Sets denote discrete candidates; intervals denote ranges. Optimization entries apply to PLARE and all pretrained baselines; downstream optimization entries cover PTB-XL. Epoch budgets are maxima; batch sizes are configured values, not effective global sizes. Filters cover PTB-XL and CPSC2018 for PLARE and baselines sharing its preprocessing. Bandpass rows list low and high cutoff candidates; dashes indicate that only downstream filters are summarized.

Parameter	Pretraining	LP	FT
Optimizer	LAMB, AdamW, Adam	AdamW	AdamW
Initial learning rate	$[10^{-5}, 10^{-3}]$	$[10^{-5}, 10^{-4}]$	$[10^{-5}, 10^{-4}]$
Weight decay	$[0, 10^{-1}]$	$[10^{-7}, 10^{-5}]$	$[10^{-7}, 10^{-5}]$
Batch size	$\{32, 64, 128, 512\}$	$\{32, 64\}$	$\{32, 64\}$
Epoch budget	200	100	100
Bandpass low cutoff (Hz)	—	$\{0.5, 1.0\}$	$\{0.5, 1.0\}$
Bandpass high cutoff (Hz)	—	$\{60, 80, 120\}$	$\{60, 80, 120\}$
Powerline removal (Hz)	—	$\{50, 60, 100, 120\}$	$\{50, 60, 100, 120\}$

Reporting conventions. Section 4.3 defines the metrics and main-text reporting conventions. Appendix captions specify the metric and comparison group used for emphasis; equal displayed values share a rank, and SD is not used to break ties. For repeated runs, we compute C from each seed's AUROC and AUPRC, then report the mean and sample SD across seeds. Component statistics are rounded independently. Residual controls retain the reported aggregate C statistics because their individual seed results were unavailable. Component metrics generally retain two decimals; complete PLARE Top-k results and internal efficiency summaries retain three. The WCR CPSC ranking exclusion is defined in Appendix B.1. Efficiency summaries average each metric across the four PTB-XL tasks; for five-seed summaries, task averaging precedes the mean and sample SD across seeds. Paired-bootstrap analyses use full-precision prediction-derived metrics. These rankings describe the reported values; they do not establish statistical significance. The external comparisons in Section 5 use fixed configurations rather than dataset-specific choices from these rankings.

B.3 Baseline Implementations and Adaptations

We reimplemented all pretrained baselines from their original papers and available public code: SimCLR (Chen et al., 2020), CLoCS (Kiyasseh et al., 2021), COMET (Wang et al., 2023), ST-MEM (Na et al., 2024), MAE (He et al., 2022), CREMA and CREMA-Max (Song et al., 2024), ECGFounder (Li et al., 2025b), ECGFM (Zhang et al., 2025b), and ECG-FM (McKeen et al., 2025). All were pretrained on the shared unlabeled ECG pool; common input processing, HPO, and LP/FT evaluation procedures are described in Appendix B.2.

Adapted foundation-model comparisons. To accommodate the shared unlabeled ECG pool and our training environment, we adapted the objectives or training configurations of the three foundation models as required. ECGFounder–CommonSSL replaced the diagnostic positive–unlabeled objective with two-view InfoNCE, while ECGFM–CPC used the contrastive predictive coding objective alone. ECG-FM–WCR followed the public pretraining example, with changes to GPU count, automatic mixed precision, validation procedures, and data eligibility rules. Public checkpoints with known pretraining overlap with downstream test data were excluded.

B.4 Paired Sampling and Training Workflow

Paired sampling. PLARE training data were divided among four sampling workers before same-patient pairing. Each worker excluded singleton patients and one ECG from each odd-sized group; recordings were not paired with themselves. Training pairs were resampled with replacement to equalize worker sample counts, without explicit corpus balancing or source weighting. Pair membership and index order were fixed at sampler initialization and reused across epochs, while augmented views varied.

Sampling scale of the representative run. The following counts describe the representative PLARE-Joint $(4, 4)$, Top-1 run; they are not measured sampling counts for the primary PLARE run. Table 7(b) gives upper bounds on unique training ECGs and patient IDs, calculated by applying the pairing rule to the full training partition before dividing it among workers. Worker partitioning could exclude further recordings; incomplete training batches were also excluded. Observed per-corpus unique counts and sampling proportions are unavailable. Each of the 200 epochs contained 4,499 batches of 64 ECGs per worker, totaling 1,151,744 ECG selections, or 575,872 pair selections, across four workers. These counts include repeated selections and precede the construction of original and augmented views.

SSL validation and selection. Each worker constructed pairs from the full validation partition, included every base pair once, and sampled additional pairs with replacement; incomplete batches were retained. For the same representative run, the worker with rank 0 processed 63,232 ECG selections per evaluation from 45,510 unique ECGs and 9,397 patient IDs. Repeated selections contributed to the SSL validation loss. The unique counts are independent of which ECG is omitted from an odd-sized group. Duplicate evaluations on other workers were not counted as additional samples. This SSL validation loss was used for checkpoint selection (Appendix B.2).

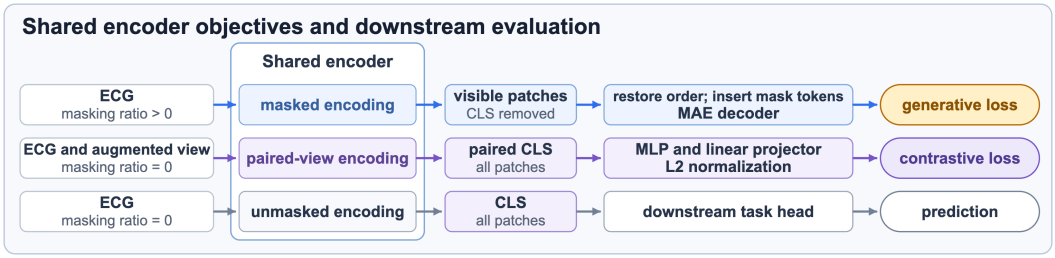

Figure 7: Shared-encoder workflow for masked reconstruction, contrastive learning, and downstream evaluation.

Self-supervised pretraining. Pretraining uses a masked forward pass on paired recordings for reconstruction and an unmasked forward pass on the recordings and their augmented views for the three contrastive relations defined in Appendix A.5. Both passes share the ECG encoder. The encoder, reconstruction decoder, and contrastive projector are optimized jointly. Routing penalties are computed within each pass and combined as specified in Appendix A.4.

Downstream evaluation. LP trains a linear classifier on the frozen encoder's final CLS representation; FT jointly updates the encoder and a dataset-specific predictor. Both use unmasked inputs and binary cross-entropy, without the reconstruction decoder or contrastive projector. FT also includes the single-pass routing penalty. LP and FT start independently from the selected pretrained encoder. Checkpoint selection follows Appendix B.2; PLARE's schedules are specified in Appendix B.5.

B.5 PLARE Training Settings and Repeated Experiments

Final settings. The implementation used capacity multiplier $\alpha_{\text{cap}} = 1.5$, noise floor $\varepsilon_{\text{noise}} = 10^{-3}$, entropy stabilizer $\varepsilon_{\text{ent}} = 10^{-6}$, load-balancing weight $\lambda_{\text{bal}} = 2$, and logit penalty weight $\lambda_z = 0.01$. The masking ratio was $\rho = 0.75$, contrastive temperature $\tau = 0.07$, and contrastive weight $\lambda_{\text{con}} = 0.005$.

Table 9 summarizes the shared HPO search space; the settings below were used for the reported PLARE experiments. Pretraining used AdamW with initial learning rate 1.8962×10^{-4} and weight decay 5.0714×10^{-4}. The batch size was 64 per worker across four workers. On PTB-XL and CPSC, LP and FT used 60 and 20 epochs, respectively, with batch size 32, initial learning rate 4×10^{-5}, and weight decay 10^{-6}. Learning rates followed linear warmup, then cosine decay (Loshchilov & Hutter, 2017) to 10^{-6}; warmup occupied 2.5% of pretraining updates and 10% of downstream updates. Router learning rates were one quarter of the base rate when the encoder was updated. Training used mixed precision and gradient-norm clipping at 1. Linear probing used a single linear output layer; the fine-tuning predictor had one hidden layer of width 1024 with dropout 0.5.

Augmented views. Pretraining constructed each augmented view with two transformations drawn from noise injection (powerline, electromyographic, and partial white noise), baseline wander or shift, smoothing and low-pass filtering, magnitude or frequency warping, angle rotation, temporal warping, crop-and-resize, time/channel/spectrogram masking, and segment permutation or concatenation. Augmentation was applied with probability 1. The implementation's intensity parameter was 8, with lead dependency disabled. This scalar identifies the configured augmentation strength, not a common physical scale across transformations.

Routing schedules. Temperature $\vartheta(t)$ and noise scale $\gamma(t)$ decreased linearly from 2 to 1 and from 1 to 0.1, respectively, while the relative confidence coefficient $\beta(t)$ increased linearly from 0 to 0.5. These transitions spanned 40 epochs in 200-epoch pretraining and 8 epochs in 20-epoch fine-tuning; values remained at their endpoints thereafter. The auxiliary coefficient $a(t)$ decreased linearly from 0.05 to 0.01 over the full training duration. Hence the effective confidence coefficient is $a(t)\beta(t)$, rather than $\beta(t)$ alone. The schedules applied to all enabled MoE modules during pretraining and fine-tuning; the encoder was frozen for linear probing.

Multi-seed experiments. Checkpoint selection followed Appendix B.2. Baseline comparisons, label-level analyses, and paired-bootstrap comparisons used models fitted with a common downstream seed. The INCART allocation analyses used frozen pretrained encoders without downstream fitting. Replacement, contextual-extension, expert-count, and Top-k comparisons used five downstream seeds, holding each configuration's selected pretrained checkpoint fixed. In the patch expert selection comparison, Top-1 entries reuse the primary PLARE runs. Sample SDs ($n = 5$, denominator $n - 1$) quantify downstream variation conditional on that checkpoint; they exclude variation from pretraining seeds and pretrained-checkpoint selection. Bootstrap intervals instead quantify sampling uncertainty conditional on the fitted models.

C Detailed Downstream Results

This appendix complements the external comparisons in Section 5.1 with label-level AUROC/AUPRC and paired bootstrap comparisons between PLARE and CREMA. These analyses use the same single-run predictions as the main comparison.

C.1 Complete Label-Level PLARE Results

Table 10 reports absolute AUROC and AUPRC for all 68 dataset-label entries (59 PTB-XL and nine CPSC) under LP and FT, using one fitted PLARE model per dataset and protocol. These analyses used test predictions without refitting or further model selection. They describe label-level variation within the single-run comparison in Section 5.1; the five-seed controlled comparisons are reported separately.

Table 10: Per-label test AUROC / AUPRC (%) for PLARE under LP and FT, using one fitted model per dataset and protocol. N_+ counts positive ECGs.

Label	N_+	LP	FT	Label	N_+	LP	FT
Super ($N = 2,162$)				**Rhythm ($N = 2,101$)**			
CD	498	90.29 / 80.79	92.18 / 85.27	**AFIB**	152	98.36 / 93.07	98.73 / 95.52
HYP	263	88.23 / 62.55	89.75 / 65.14	**AFLT**	7	96.62 / 61.38	95.77 / 73.69
MI	553	91.16 / 80.67	93.45 / 84.94	**BIGU**	8	91.92 / 23.30	98.19 / 35.71
NORM	963	93.26 / 90.56	94.32 / 92.15	**PACE**	28	96.24 / 78.44	97.90 / 83.61
STTC	523	92.25 / 80.13	93.59 / 82.57	**PSVT**	2	99.29 / 53.12	99.88 / 64.29
				SARRH	77	78.61 / 13.80	96.78 / 62.80
Sub ($N = 2,162$)				**SBRAD**	64	94.58 / 57.74	96.50 / 72.85
AMI	308	96.08 / 83.77	96.99 / 85.56	**SR**	1677	87.96 / 95.65	96.38 / 98.86
CLBBB	54	99.92 / 96.89	99.94 / 97.43	**STACH**	82	99.15 / 81.17	99.29 / 92.86
CRBBB	54	99.79 / 90.90	99.83 / 91.85	**SVARR**	14	85.04 / 4.26	92.95 / 18.47
ILBBB	8	90.98 / 13.54	91.31 / 28.93	**SVTAC**	3	99.54 / 21.44	99.81 / 33.13
IMI	328	92.81 / 71.53	94.26 / 78.42	**TRIGU**	2	94.35 / 50.42	98.86 / 5.98
IRBBB	112	97.11 / 67.20	97.67 / 77.15				
ISCA	94	90.41 / 29.60	93.55 / 37.79	**Form ($N = 882$)**			
ISCI	40	91.43 / 33.42	95.03 / 42.30	**ABQRS**	322	79.88 / 66.85	84.95 / 76.64
ISC_	128	95.34 / 64.47	97.15 / 72.04	**DIG**	18	88.83 / 17.20	86.64 / 14.31
IVCD	79	75.22 / 14.78	76.44 / 16.83	**HVOLT**	6	94.50 / 9.23	95.22 / 22.85
LAFB/LPFB	180	96.42 / 76.23	97.67 / 83.78	**INVT**	29	84.00 / 18.37	94.86 / 48.57
LAO/LAE	42	87.03 / 18.93	88.17 / 16.85	**LNGQT**	11	86.46 / 10.14	87.22 / 15.85
LMI	20	88.34 / 6.35	92.29 / 9.07	**LOWT**	44	71.33 / 11.41	84.83 / 22.89
LVH	214	94.07 / 70.05	93.98 / 72.58	**LPR**	34	94.61 / 43.32	95.36 / 57.77
NORM	963	93.14 / 90.52	94.36 / 91.06	**LVOLT**	18	92.72 / 28.62	89.87 / 21.61
NST_	77	81.93 / 14.62	85.14 / 17.24	**NDT**	183	85.69 / 58.51	88.77 / 63.68
PMI	2	81.40 / 50.12	99.69 / 31.67	**NST_**	77	73.24 / 25.63	74.87 / 27.53
RAO/RAE	10	98.63 / 46.03	97.27 / 51.13	**NT_**	42	75.46 / 12.30	88.25 / 26.72
RVH	12	97.89 / 22.68	95.16 / 19.20	**PAC**	40	74.06 / 13.73	95.30 / 83.38
SEHYP	3	99.86 / 64.65	98.33 / 43.80	**PRC(S)**	1	77.58 / 0.50	75.77 / 0.47
STTC	223	86.07 / 41.90	90.33 / 51.92	**PVC**	115	98.30 / 88.53	99.44 / 94.29
WPW	8	95.62 / 59.59	87.64 / 62.20	**QWAVE**	55	76.16 / 21.82	78.94 / 26.06
_AVB	83	96.88 / 57.50	98.03 / 66.75	**STD_**	101	78.41 / 32.53	83.73 / 41.69
				STE_	3	70.29 / 0.73	81.85 / 1.20
CPSC ($N = 1,291$)				**TAB_**	3	76.58 / 1.07	83.50 / 2.39
NORM	183	93.67 / 74.78	97.23 / 84.41	**VCLVH**	87	84.98 / 37.14	89.25 / 50.14
AF	197	98.38 / 95.33	98.71 / 96.68				
I-AVB	137	98.39 / 91.65	99.14 / 96.36				
LBBB	36	99.48 / 91.88	98.32 / 91.98				
RBBB	316	98.09 / 94.33	98.94 / 96.72				
PAC	106	79.06 / 31.16	95.32 / 76.89				
PVC	122	92.79 / 74.19	95.62 / 85.24				
STD	158	95.25 / 79.49	96.24 / 86.27				
STE	36	92.88 / 44.92	97.64 / 64.51				

C.2 PAIRED PLARE–CREMA COMPARISONS

Tables 12 and 13 report PLARE's per-label differences from CREMA for all 12 Rhythm and nine CPSC labels. Table 11 summarizes the all-label macro-averaged differences. These supplementary analyses use C as defined in Section 4.3, with AUPRC computed as noninterpolated average precision. A positive ΔC does not establish improvement in both components, and its interval is not an interval for either component separately. The baseline comparisons evaluate complete models with different architectures and training procedures (Appendix B.3).

Paired resampling. Predictions were matched by recording and window identifiers with identical label vectors, and LP and FT were analyzed separately. Rhythm comprised 2,101 test ECGs: 2,098 from 1,823 identified patients and three without patient mappings. A patient-cluster bootstrap (Field & Welsh, 2007) sampled 1,826 clusters, retaining all ECGs of each identified patient and treating

each unmatched ECG as a separate cluster. CPSC used recordings as the sampling units for the 1,291 test ECGs in the common 7:1:2 split because patient identities were unavailable.

Each analysis used 1,000 paired draws shared across models and labels, with metrics and differences recomputed within each draw. The 95% percentile intervals are pointwise, unadjusted, and conditional on the fitted models; they describe sampling uncertainty, not training-seed variation. A draw with only one class was excluded for that label; a macro-average draw was excluded if any label was unevaluable. Macro intervals used 722 LP and 702 FT draws for all Rhythm labels, and 1,000 for CPSC. They were computed from metrics computed within each paired bootstrap resample, not by averaging label-interval endpoints. All estimates used full-precision predictions. The direct macro interval for CPSC LP PLARE versus CREMA was unavailable; Table 11 retains the point estimate.

Table 11: Macro-averaged ΔC [95% paired CI] for PLARE relative to CREMA, in percentage points. All-label summaries include 12 Rhythm or nine CPSC labels. A dash within brackets denotes an unavailable CI.

Dataset	**LP: All-label ΔC [95% CI]**	**FT: All-label ΔC [95% CI]**
Rhythm	$+7.606$ $[+2.700, +11.491]$	$+7.159$ $[+4.652, +9.296]$
CPSC	$+3.567$ [—]	$+3.603$ $[+2.153, +5.002]$

The label-level tables identify labels with higher or lower performance relative to CREMA. They do not establish a causal link between expert allocation and downstream performance.

Table 12: All 12 Rhythm labels for PLARE versus CREMA: ΔC_ℓ [pointwise 95% paired CI], in percentage points. N_+: positive ECGs; B: valid draws out of 1,000, for the PLARE–CREMA contrast.

Label	N_+	**LP**		**FT**	
		B	**ΔC_ℓ [95% CI]**	B	**ΔC_ℓ [95% CI]**
AFIB	152	1,000	$+5.622$ $[+2.614, +8.899]$	1,000	$+1.992$ $[+0.356, +3.758]$
AFLT	7	996	$+9.958$ $[-17.466, +30.696]$	997	$+6.502$ $[0.000, +13.106]$
BIGU	8	1,000	$+3.936$ $[-8.899, +16.765]$	1,000	$+13.843$ $[+1.773, +26.209]$
PACE	28	1,000	$+5.143$ $[+0.827, +10.359]$	1,000	$+5.085$ $[+0.591, +12.124]$
PSVT	2	876	$+12.241$ $[-1.769, +33.381]$	861	-2.881 $[-19.112, +3.577]$
SARRH	77	1,000	$+4.567$ $[-0.022, +8.889]$	1,000	$+25.001$ $[+17.468, +32.162]$
SBRAD	64	1,000	-1.917 $[-7.312, +3.323]$	1,000	$+9.609$ $[+4.800, +15.029]$
SR	1,677	1,000	-0.605 $[-1.669, +0.441]$	1,000	$+4.362$ $[+3.197, +5.704]$
STACH	82	1,000	$+4.240$ $[+0.386, +8.124]$	1,000	$+3.685$ $[+0.570, +7.034]$
SVARR	14	1,000	$+3.904$ $[-2.407, +11.705]$	1,000	$+8.781$ $[-3.155, +18.878]$
SVTAC	3	952	$+9.826$ $[+2.317, +34.421]$	944	$+6.611$ $[-3.213, +23.054]$
TRIGU	2	869	$+34.359$ $[+9.601, +58.789]$	868	$+3.314$ $[-0.884, +11.051]$

Table 13: All nine CPSC labels for PLARE versus CREMA: ΔC_ℓ [pointwise 95% paired CI], in percentage points. N_+: positive ECGs; B: valid draws out of 1,000, for the PLARE–CREMA contrast.

Label	N_+	**LP**		**FT**	
		B	**ΔC_ℓ [95% CI]**	B	**ΔC_ℓ [95% CI]**
NORM	183	1,000	$+1.285$ $[-1.362, +3.913]$	1,000	-0.055 $[-2.202, +2.035]$
AF	197	1,000	$+2.307$ $[+1.212, +3.661]$	1,000	$+1.812$ $[+0.914, +2.769]$
I-AVB	137	1,000	$+7.565$ $[+4.498, +10.526]$	1,000	$+2.391$ $[+0.770, +4.197]$
LBBB	36	1,000	$+0.084$ $[-3.486, +3.499]$	1,000	$+5.548$ $[+0.846, +12.040]$
RBBB	316	1,000	$+0.842$ $[-0.055, +1.971]$	1,000	$+1.277$ $[+0.394, +2.365]$
PAC	106	1,000	$+5.926$ $[+0.076, +11.485]$	1,000	$+14.787$ $[+10.016, +19.450]$
PVC	122	1,000	$+8.651$ $[+5.012, +12.429]$	1,000	$+0.539$ $[-2.197, +3.286]$
STD	158	1,000	$+4.657$ $[+1.769, +7.636]$	1,000	$+1.734$ $[-0.633, +4.216]$
STE	36	1,000	$+0.783$ $[-5.604, +7.121]$	1,000	$+4.390$ $[-3.118, +10.694]$

D ABLATION DETAILS

The following results complement Section 5.2, in the same order: feature integration, content-dependent routing, and the number of selected patch experts. For these comparisons, we report AUROC and AUPRC alongside the C summaries presented in the main text.

D.1 FEATURE REPLACEMENT AND CONTENT-DEPENDENT ROUTING

This subsection details the two component ablations in Section 5.2. Table 14 compares PLARE, w/o Repl. (PLARE-Residual), and w/o CDR in the same order as Tables 3 and 4, reporting AUROC, AUPRC, and C with their sample SDs.

Table 14: Feature replacement and routing ablations corresponding to Tables 3 and 4: AUROC, AUPRC, and C on the 0–100 scale (mean $\pm$ sample SD, five downstream seeds). Repl.: replacement; CDR: content-dependent routing. Component and C summaries retain their reported precision.

Metric	Model	Super	Sub	Rhythm	Form	CPSC
Linear probing (LP)						
AUROC	**PLARE**	91.08 ± 0.02	92.65 ± 0.36	93.89 ± 0.33	79.39 ± 1.98	94.23 ± 0.13
	w/o Repl.	91.36 ± 0.05	92.67 ± 0.17	91.53 ± 0.38	79.09 ± 1.45	94.23 ± 0.35
	w/o CDR	90.010 ± 0.046	92.738 ± 0.109	93.442 ± 0.463	79.022 ± 0.934	94.140 ± 0.110
AUPRC	**PLARE**	78.90 ± 0.09	51.76 ± 0.87	49.63 ± 1.92	25.76 ± 0.59	75.18 ± 0.12
	w/o Repl.	79.20 ± 0.26	49.26 ± 1.26	47.66 ± 1.80	25.91 ± 0.43	74.46 ± 0.59
	w/o CDR	77.496 ± 0.042	50.976 ± 0.983	46.152 ± 1.403	24.830 ± 0.400	73.220 ± 0.291
C	**PLARE**	84.99 ± 0.05	72.21 ± 0.55	71.76 ± 0.83	52.57 ± 1.06	84.70 ± 0.08
	w/o Repl.	85.28 ± 0.15	70.97 ± 0.59	69.59 ± 0.89	52.50 ± 0.73	84.35 ± 0.46
	w/o CDR	83.753 ± 0.010	71.857 ± 0.500	69.797 ± 0.809	51.926 ± 0.538	83.680 ± 0.182
Fine-tuning (FT)						
AUROC	**PLARE**	92.66 ± 0.21	93.38 ± 0.40	96.92 ± 0.46	86.04 ± 1.33	97.41 ± 0.14
	w/o Repl.	92.50 ± 0.16	92.68 ± 0.66	95.74 ± 0.50	85.82 ± 1.25	97.52 ± 0.12
	w/o CDR	92.054 ± 0.195	92.152 ± 0.797	94.284 ± 1.012	86.790 ± 1.253	97.082 ± 0.169
AUPRC	**PLARE**	81.96 ± 0.33	52.73 ± 1.49	61.35 ± 1.12	36.59 ± 0.65	86.04 ± 0.58
	w/o Repl.	81.60 ± 0.59	51.48 ± 1.30	61.88 ± 0.85	36.67 ± 0.42	86.13 ± 0.36
	w/o CDR	81.544 ± 0.265	51.930 ± 0.661	58.844 ± 1.152	35.190 ± 0.530	85.056 ± 0.366
C	**PLARE**	87.31 ± 0.27	73.05 ± 0.86	79.14 ± 0.66	61.32 ± 0.94	91.72 ± 0.34
	w/o Repl.	87.05 ± 0.36	72.08 ± 0.57	78.81 ± 0.23	61.25 ± 0.47	91.83 ± 0.24
	w/o CDR	86.799 ± 0.217	72.041 ± 0.591	76.564 ± 0.971	60.990 ± 0.846	91.069 ± 0.260

Replacement versus residual integration. PLARE and PLARE-Residual used $(P, S) = (4, 0)$ with Top-1 selection. Only the additive residual connection varied between them. For PLARE-Residual, the feature supplied to token projection is

$$\overline{\mathbf{u}}_p^{\mathrm{res}} = \mathbf{u}_p + \alpha_{\mathrm{patch}} \mathcal{R}_{\mathrm{patch}}(\mathbf{h}_p^{\mathrm{patch}}), \tag{19}$$

where $\alpha_{\mathrm{patch}} = 1$ was fixed throughout training in both configurations and $\mathbf{h}_p^{\mathrm{patch}} = \mathrm{LN}_{\mathrm{patch,in}}(\mathbf{u}_p)$. The shared branch uses the unnormalized convolutional feature. Replacement and residual addition share the output normalization, linear projection, and RMS normalization in Equation 6. If no expert assignment is retained after capacity filtering, $\mathcal{R}_{\mathrm{patch}} = \mathbf{0}$, so both variants pass $\mathbf{u}_p$ to this common projection path.

Content-dependent routing. The w/o CDR control uses a fixed mapping from patch position within the input to expert index. For the 82 positions in each ten-second input, we generated this mapping once using a dedicated random seed of 20260921, with 21, 21, 20, and 20 positions assigned to the four experts. The mapping is stored as `position_assignment` in the checkpoint and reused for every ECG throughout training, validation, and testing, remaining fixed across all epochs. Assignments depend only on patch position, independently of patient or record identity and waveform content. Specialization by temporal position remains possible.

Checkpoint selection and reporting. For PLARE and PLARE-Residual, we selected one pretrained checkpoint by the lowest validation SSL loss over 200 epochs for each variant and held it fixed across the five downstream seeds. For each downstream seed, we selected the downstream checkpoint by the highest validation C.

The residual results were available as aggregate means and sample SDs, including the mean and sample SD of seed-wise C, but individual downstream-seed values were unavailable for paired analysis. The component results complement the C-based interpretation in Section 5.2; rankings of the reported means do not establish statistical significance.

D.2 NUMBER OF SELECTED PATCH EXPERTS

Table 15 reports the AUROC and AUPRC components of the PLARE comparison in Figure 4. We varied $k_P \in \{1, 2, 4\}$ with $P = 4$ and no contextual experts. Each setting was pretrained separately and used the same selection count during downstream evaluation. Results summarize five downstream seeds for each setting's fixed pretrained checkpoint; the $k_P = 1$ entries reuse the primary PLARE runs.

Selecting all experts preserves the expert architectures and count, as well as auxiliary routing losses, while uniformly combining all expert outputs (Appendix A.3). Thus, these comparisons evaluate both the learned representations and the number of expert transformations at inference. The sample SDs quantify downstream variability conditional on each pretrained checkpoint.

By C, Top-1 ranked highest on Sub, Rhythm, and CPSC under LP and on Super, Sub, and Rhythm under FT. The highest values on the remaining tasks were obtained with $k_P = 2$ on Super and $k_P = 4$ on Form under LP, and with $k_P = 2$ on Form and CPSC under FT. The component results show why AUROC rankings alone need not match the C-based comparison: on CPSC under LP, $k_P = 4$ yielded 0.060 points higher AUROC but 1.168 points lower AUPRC than $k_P = 1$.

Table 15: Component results for the number of selected patch experts in PLARE ($P = 4$, $S = 0$): AUROC and AUPRC (%; mean ± sample SD over five downstream seeds). Both component cells are bold/underlined for the best/second-ranked settings by C, computed from the displayed component means within each (P, S), dataset, and protocol; ties share ranks. A dash denotes the absence of contextual experts.

				LP		FT	
(P, S)	Dataset	k_P	k_S	AUROC	AUPRC	AUROC	AUPRC
$(4, 0)$	**Super**	1	—	91.080 ± 0.023	78.896 ± 0.094	**92.664 ± 0.208**	**81.964 ± 0.331**
$(4, 0)$	**Super**	2	—	**91.386 ± 0.059**	**79.548 ± 0.160**	92.428 ± 0.190	81.594 ± 0.215
$(4, 0)$	**Super**	4	—	91.000 ± 0.058	78.862 ± 0.141	92.686 ± 0.271	81.864 ± 0.519
$(4, 0)$	**Sub**	1	—	**92.654 ± 0.360**	**51.758 ± 0.865**	**93.376 ± 0.403**	**52.726 ± 1.489**
$(4, 0)$	**Sub**	2	—	92.782 ± 0.218	49.578 ± 1.106	93.222 ± 0.147	52.662 ± 1.073
$(4, 0)$	**Sub**	4	—	92.618 ± 0.217	49.048 ± 1.019	92.534 ± 0.378	51.002 ± 1.497
$(4, 0)$	**Rhythm**	1	—	**93.894 ± 0.329**	**49.634 ± 1.917**	**96.920 ± 0.464**	**61.352 ± 1.121**
$(4, 0)$	**Rhythm**	2	—	92.400 ± 0.229	44.916 ± 1.060	96.074 ± 0.766	61.968 ± 1.908
$(4, 0)$	**Rhythm**	4	—	92.906 ± 0.467	48.182 ± 1.155	95.992 ± 0.622	59.412 ± 1.052
$(4, 0)$	**Form**	1	—	79.386 ± 1.983	25.760 ± 0.585	86.040 ± 1.325	36.592 ± 0.655
$(4, 0)$	**Form**	2	—	77.180 ± 0.751	25.288 ± 0.794	**87.320 ± 1.928**	**36.624 ± 0.968**
$(4, 0)$	**Form**	4	—	**80.546 ± 1.859**	**26.848 ± 0.498**	86.570 ± 0.557	36.364 ± 1.579
$(4, 0)$	**CPSC**	1	—	**94.226 ± 0.125**	**75.184 ± 0.123**	97.408 ± 0.137	86.040 ± 0.584
$(4, 0)$	**CPSC**	2	—	93.524 ± 0.255	72.652 ± 0.670	**97.454 ± 0.193**	**86.478 ± 0.616**
$(4, 0)$	**CPSC**	4	—	94.286 ± 0.258	74.016 ± 0.383	97.278 ± 0.106	85.614 ± 0.719

E CONTEXTUAL EXPERT EXTENSION DETAILS

This comparison complements Section 5.5 by reporting AUROC, AUPRC, and C for PLARE $(P, S) = (4, 0)$ and PLARE-Joint $(4, 4)$. Both use four patch experts with Top-1 selection. PLARE-Joint replaces the standard FFNs in Transformer blocks B4–B6 with contextual MoE modules, each containing $S = 4$ experts and using Top-1 selection, as specified in Appendix A.6. Each configuration used five downstream seeds with its selected pretrained checkpoint fixed, following Appendix B.5.

Table 16: Contextual extension results under LP and FT: AUROC (%), AUPRC (%), and C (0–100), reported as mean ± sample SD across five downstream seeds with a fixed pretrained checkpoint per configuration. Both configurations use Top-1 selection. Bold marks the higher displayed mean in each pair.

Panel	Configuration	Super	Sub	Rhythm	Form	CPSC
LP AUROC	**PLARE-Joint**	**91.20** ± 0.05	92.52 ± 0.25	92.20 ± 0.55	79.10 ± 1.09	94.17 ± 0.17
	PLARE	91.08 ± 0.02	**92.65** ± 0.36	**93.89** ± 0.33	**79.39** ± 1.98	**94.23** ± 0.13
LP AUPRC	**PLARE-Joint**	**79.56** ± 0.05	50.48 ± 1.09	49.14 ± 1.33	**26.35** ± 0.21	74.71 ± 0.14
	PLARE	78.90 ± 0.09	**51.76** ± 0.87	**49.63** ± 1.92	25.76 ± 0.59	**75.18** ± 0.12
LP C	**PLARE-Joint**	**85.38** ± 0.05	71.50 ± 0.58	70.67 ± 0.60	**52.73** ± 0.48	84.44 ± 0.13
	PLARE	84.99 ± 0.05	**72.21** ± 0.55	**71.76** ± 0.83	52.57 ± 1.06	**84.70** ± 0.08
FT AUROC	**PLARE-Joint**	92.58 ± 0.15	93.32 ± 0.23	96.47 ± 0.82	**86.32** ± 1.14	97.40 ± 0.08
	PLARE	**92.66** ± 0.21	**93.38** ± 0.40	**96.92** ± 0.46	86.04 ± 1.33	**97.41** ± 0.14
FT AUPRC	**PLARE-Joint**	81.88 ± 0.31	51.15 ± 1.18	61.13 ± 1.80	36.10 ± 0.97	**86.20** ± 0.33
	PLARE	**81.96** ± 0.33	**52.73** ± 1.49	**61.35** ± 1.12	**36.59** ± 0.65	86.04 ± 0.58
FT C	**PLARE-Joint**	87.23 ± 0.23	72.24 ± 0.61	78.80 ± 1.00	61.21 ± 0.72	**91.80** ± 0.20
	PLARE	**87.31** ± 0.27	**73.05** ± 0.86	**79.14** ± 0.66	**61.32** ± 0.94	91.72 ± 0.34

Table 16 shows that PLARE achieved higher mean C on three of five tasks under LP and four of five under FT. PLARE-Joint exceeded PLARE only on Super and Form under LP, by 0.39 and 0.16 points, respectively, and on CPSC under FT, by 0.08 points. Thus, adding contextual experts did not consistently improve downstream performance over PLARE under the evaluated Top-1 setting.

F INFERENCE COST AND LP PERFORMANCE

This appendix documents the measurements supporting the performance–cost comparison in Section 5.3. Costs in Table 17 include the encoder and its linear prediction head. We profiled 12-lead inputs on an A100-SXM4-40GB in FP32 at batch size one. Each of ten processes measured median latency over 600 forward passes and peak memory. We report the mean ± SD of these process-level measurements; memory SDs all round to zero at four decimal places and are omitted. The external variants retained their pretraining input sampling rates and durations.

Active (selected-path) FLOPs count shared computation and retained expert transformations; values for PLARE and its expert variants are mean ± SD over 32 ECGs. Total FLOPs count shared computation and all expert transformations, providing a reference for computation with all experts. The two counts coincide for the baselines and all-expert selection. Latency and memory describe the profiled execution, which excluded auxiliary computations and reused identical routing results.

Table 17: Inference costs for the main comparisons. PLARE uses 250 Hz, 10 s inputs and Top-1 patch expert selection. Active FLOPs count selected expert transformations; total FLOPs include all expert transformations.

(a) Baseline costs

Model	Input (Hz, s)	Params (M)	Total GFLOPs	Latency (ms)	Peak memory (GB)
SimCLR	250, 10	18.041	3.518	2.187 ± 0.017	0.0883
CLoCS	250, 10	28.542	5.633	2.703 ± 0.013	0.1383
COMET	250, 10	28.542	5.633	2.712 ± 0.014	0.1383
ST-MEM	250, 10	26.540	13.901	2.083 ± 0.005	0.1367
MAE	250, 10	18.041	3.518	2.187 ± 0.010	0.0883
CREMA	250, 10	18.041	3.518	2.190 ± 0.010	0.0883
CREMA-Max	250, 10	177.601	30.586	6.917 ± 0.039	0.7484
ECGFounder–CommonSSL	500, 10	31.982	2.337	8.670 ± 0.059	0.1407
ECGFM–CPC	500, 10	20.058	1.980	2.150 ± 0.008	0.0959
ECG-FM–WCR	500, 5	90.887	29.303	5.452 ± 0.007	0.4106

Table 17: Inference costs for the main comparisons (continued).

(b) Primary PLARE cost

Configuration	Params (M)	Active GFLOPs	Total GFLOPs	Latency (ms)	Peak memory (GB)
PLARE (4,0)	69.528	8.686 ± 0.012	12.318	4.611 ± 0.042	0.2959

Performance aggregation. The main comparison in Table 5 uses one run per model. Its AUROC/AUPRC means were 87.92/47.44 for CREMA, 87.97/48.40 for CREMA-Max, 84.32/45.00 for ECG-FM–WCR, and 89.81/52.37 for PLARE. These single-run metrics were averaged separately across the four PTB-XL tasks and rounded to two decimals before computing C.

G EXPERT ALLOCATION IN PLARE

This appendix supplements Section 5.4 with implementation details, inclusion rules, and probability-based results.

Input processing and capacity. INCART was used only for analysis: no model training, checkpoint selection, or downstream head fitting was performed on these data. We analyzed recordings from 32 patients, dividing each thirty-minute recording into nonoverlapping ten-second windows (Yakushenko, 2008; Pollard et al., 2026). Signals were resampled from 257 to 250 Hz using the preprocessing configuration from SSL pretraining (lead filling, resampling, and length adjustment). No additional filtering, normalization, augmentation, or masking was used. Evaluation used batch size one, routing temperature one, and capacity factor 1.5, giving a capacity of 31 assignments per expert.

Capacity filtering rejected 16.056% of selections during the allocation analysis. Rejected patches retained their pre-expert features; the tokens were not removed. A pre-capacity selection therefore need not imply that the corresponding expert transformation contributed to the token. This rejection rate does not estimate rejection during training.

Beat inclusion and probability aggregation. We also applied the beat-context averaging in Section 5.4 to router probability vectors. Both neighboring annotations and the full padded interval had to lie in the original input's token-covered context; intervals were not clipped for analysis. Overlapping intervals could share patches. INCART beat annotations locate beats rather than delineating P/QRS/T boundaries, so these summaries measure beat context rather than isolated waveform components. Only recordings containing both N and V beats were retained, contributing 127,203 N and 18,024 V beats.

Statistical aggregation and uncertainty. Record-level base-2 JS divergences were averaged equally within each patient and then equally across patients. This weighting prevents patients with more recordings or beats from dominating the result. In the within-window permutations, beat labels other than N/V were also held fixed. The reference assumes exchangeability of N/V labels within each window. Patient-bootstrap intervals used 10,000 resamples. Holm correction covered the two primary tests, using probabilities and Top-1 selections. These are related measurements of the same router. Table 18 reports the probability-based analysis alongside the Top-1 results.

Table 18: Normal/PVC allocation in frozen PLARE: patient-averaged JS divergence and ΔJS, in bits. Brackets give 95% patient-bootstrap intervals. Statistics use pre-capacity routing.

Measure	Observed JS	Shuffle mean	ΔJS	Holm p
Probabilities	0.027909 $[0.014721, 0.045145]$	0.016616	0.011293 $[0.005192, 0.019690]$	0.000200
Top-1	0.028103 $[0.014803, 0.045553]$	0.016739	0.011363 $[0.005185, 0.019955]$	0.000200

Scope. These statistics do not quantify rare-pattern preservation or establish a causal explanation for downstream performance. All measurements are conditional on one pretrained checkpoint.