FORGE-CTG: OUTCOME QUERIES AND DUAL-VIEW DISTILLATION FOR INTRAPARTUM FETAL ACIDOSIS DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Intrapartum cardiotocography (CTG) is used for fetal surveillance, yet detecting neonatal acidaemia from fetal heart-rate and uterine-contraction traces remains difficult. A pathological trace can trigger intervention without acidaemia, while electrode loss, motion artefact, or prolonged gaps can obscure a trace. Existing learning pipelines face limitations. They reduce a continuous physiological outcome to a **single pH threshold**, discarding severity structure, and treat missing signal as noise to be imputed rather than as a quality condition that should change supervision. To bridge this gap, we introduce FORGE-CTG, a framework built around complementary ideas. SCOS supervises a shared representation through *binary risk*, *rank-consistent ordinal pH thresholds*, continuous pH, and masked base deficit. QVD derives reliability from the raw trace and validity mask, then trains a student against an exponential-moving-average teacher reading a conservative repair; **uniform patch weighting** keeps quality in the teacher's view without weakening the student's learning signal. On the **CTU-UHB dataset**, FORGE-CTG ranks first among 19 models at **0.7283 record-level AUROC** and degrades **4.6× more slowly** under increasing missingness, reaching ranking performance with half the training labels. The severity-supervision principle also transfers to clinical time-series tasks, with **PTB-XL dataset** and **CPSC2018 dataset** experiments showing a **24%** reduction in ordinal severity error with **perfect rank consistency**. Together, these results support a practical direction for clinical monitoring in which richer supervision and explicit responses to signal quality improve the reliability of risk estimates without requiring larger models.

1 INTRODUCTION

Intrapartum cardiotocography (CTG) records fetal heart rate (FHR) and uterine contractions (UC) to support fetal surveillance during labour. Predicting neonatal acidaemia from these traces is difficult because the monitored signal and the outcome describe different aspects of physiology. A concerning trace may precede intervention without an acidotic outcome, while electrode loss and motion artefacts can obscure the signal used for prediction. The public CTU-UHB database pairs CTG with umbilical-artery pH (Chudáček et al., 2014), and recent CTG pretraining work uses its 552 records as an outcome-labelled benchmark (Fridman & Ben Shachar, 2026). The learning objective is to rank records by outcome risk while retaining useful information about severity and signal quality.

Two properties of this task motivate our approach. First, a **single pH threshold** compresses a continuous outcome into a binary label. Rank-consistent ordinal models preserve ordered categories through related binary decisions (Cao et al., 2020), but CTG supervision can also use continuous metabolic outcomes. Second, missing segments are often repaired before prediction. Missingness-aware recurrent models (Che et al., 2018) and attention-based imputation (Du et al., 2023) model incomplete time series explicitly, yet repaired values can still obscure which measurements were observed. These choices discard information at opposite ends of the pipeline: outcome structure at supervision and observation quality at input.

Our benchmark comparison suggests that both deserve attention. On 551 CTU-UHB records, including 105 positives at pH < 7.15, a clinical-feature logistic regression reaches 0.7099 AUROC

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.

and remains competitive with the strongest neural baselines. Performance varies substantially across 19 configurations, including patched transformers (Nie et al., 2023), inverted attention (Liu et al., 2024), and medical multi-granularity patching (Wang et al., 2024). This motivates a controlled study of supervision and training views on a shared backbone, rather than attributing differences to model size alone.

We introduce FORGE-CTG, which combines **severity supervision** with **quality-conditioned distillation**. Its Severity-Consistent Outcome Spectrum (SCOS) adapts outcome queries used for medical time-series classification (Huang et al., 2026) to predict binary risk, nested pH thresholds, continuous pH, and base deficit. Its Quality-Conditioned Dual-View Distillation (QVD) follows the weight-averaged teacher principle (Tarvainen & Valpola, 2017): a student reads the raw trace while an EMA teacher reads a conservatively repaired FHR trace. Signal-derived reliability enters the representation, every FHR patch contributes equally to consistency, and only the raw-input student is retained for inference. Figure 1 illustrates this distinction.

Contribution. Our contributions are threefold:

- **Severity-consistent outcome supervision.** We introduce SCOS, which represents fetal condition through complementary queries for binary acidaemia risk, rank-consistent pH thresholds, continuous pH, and masked base deficit. This outcome spectrum preserves clinically meaningful ordering and metabolic variation that would otherwise be collapsed by supervision at a single diagnostic cutoff.
- **Quality-conditioned dual-view distillation.** We propose QVD, in which a raw-view student is guided by an EMA teacher that reads a conservatively repaired FHR view. Signal-derived reliability conditions the representation, while uniform patch consistency and raw-only inference prevent repaired values from replacing the observations available to the deployed model.
- **Mechanism-focused empirical evaluation.** We benchmark 19 configurations and use backbone-matched ablations to separate the effects of richer outcome supervision and paired training views. Missingness stress tests, label-efficiency and subgroup analyses, and transfer to two ECG datasets further delineate where the proposed principles help and where their gains remain limited.

The five-seed model achieves 0.7283 record-level AUROC and 0.4226 AUPRC, ranking first among the evaluated configurations, with overlapping confidence intervals against the strongest alternatives. On the same backbone, adding SCOS and then QVD increases AUROC from 0.7028 to 0.7200 and 0.7342 in the three-seed ablation. Under random missingness, QVD reduces the magnitude of the degradation slope by $4.6\times$ relative to SCOS alone. Additional experiments examine label efficiency, signal-defined subgroups, and transfer to adult ECG. Together, these analyses identify where richer supervision and paired training views help, and where their benefits remain limited.

2 RELATED WORK

Learning from physiological time series. CTU-UHB provides paired intrapartum traces and clinical outcomes for studying CTG prediction (Chudáček et al., 2014). Time-series learning offers several relevant model families: patch-based forecasting with PatchTST (Nie et al., 2023), variable-token attention with iTransformer (Liu et al., 2024), temporal variation modelling with TimesNet (Wu et al., 2023), convolutional classification with InceptionTime (Ismail Fawaz et al., 2020), and efficient kernel transforms with MiniRocket (Dempster et al., 2021). Medical time-series transformers and CTG-specific pretraining further adapt these ideas to physiological signals (Wang et al., 2024; Fridman & Ben Shachar, 2026). Our focus is **the supervision signal**: we examine how outcome structure and paired views affect a compact encoder under a common record-level protocol.

Ordinal and multi-outcome supervision. Rank-consistent ordinal regression models ordered categories through related binary decisions (Cao et al., 2020). For pH, nested thresholds retain distinctions that one acidaemia cutoff omits. Continuous regression supplies outcome magnitude, while an auxiliary base-deficit target supplies a related metabolic measurement when available. Multi-task frameworks such as LibMTL organise methods for sharing representations and balancing objectives (Lin & Zhang, 2023). SCOS uses an equal-weight sum of binary, ordinal, and continuous losses; its contribution lies in the **outcome spectrum** exposed to the shared representation, rather than a new generic loss-balancing algorithm.

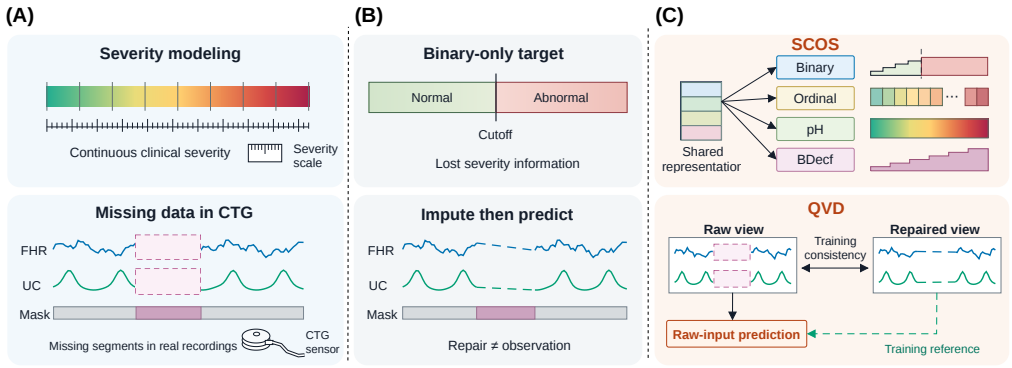

Figure 1: **Two limitations of the standard CTG pipeline and the corresponding mechanisms in FORGE-CTG.** (A) CTG outcomes vary continuously in severity, and real recordings contain missing FHR and UC segments. (B) Binary-only supervision discards severity structure, while impute-then-predict obscures which values were observed. (C) SCOS supervises a shared representation with binary risk, nested ordinal thresholds, continuous pH, and masked base deficit. QVD trains the raw-view student against an EMA teacher that reads a conservatively repaired view; the repaired view provides a training reference, while prediction uses the raw input.

Missing observations and consistency learning. GRU-D incorporates missingness and time intervals into recurrent learning (Che et al., 2018); mTAN uses attention over irregular observation times (Shukla & Marlin, 2021); Raindrop models sensor relationships (Zhang et al., 2022); and SAITS reconstructs missing values with self-attention (Du et al., 2023). Reversible instance normalisation addresses a different source of variation, changes in per-instance statistics (Kim et al., 2022). EMA-teacher methods stabilise consistency targets during training (Tarvainen & Valpola, 2017). QVD combines these concerns through **asymmetric input views**: the teacher receives a bounded repair, while the student retains the raw input. Uniform patch weighting keeps the learning signal active in the regions where observation quality is low. This differs from making an imputed trace the test-time input or down-weighting the consistency loss wherever the signal is unreliable.

3 METHOD

3.1 PREDICTION TASK AND INPUT REPRESENTATION

A CTG record consists of $x \in \mathbb{R}^{2\times T}$, containing FHR and UC sampled at 4 Hz, and a validity mask m. We use the final 30 minutes before delivery, giving $T = 7200$, and predict $y = \mathbb{1}[\text{pH} < 7.15]$. Continuous pH, four nested threshold labels, and available base-deficit measurements are training targets. **Outcome variables** are excluded from the inputs. Normalisation statistics are estimated using only the training portion of each fold.

The shared encoder has approximately 0.7M parameters, three transformer layers, hidden dimension $d = 128$, and four attention heads. Non-overlapping patches of 32 samples yield 225 tokens per channel and 450 tokens per record. A linear patch projection is combined with learned channel and positional embeddings. A second projection embeds four patch statistics: validity fraction, mean, standard deviation, and reliability. FHR statistics distinguish observed from missing samples; UC zero values remain valid. The FHR reliability feature is defined in Section 3.3; the UC reliability slot is zero in the implemented backbone. This representation exposes **signal quality** alongside signal content.

An exploratory cross-channel response branch reduced AUROC from 0.7342 to 0.7033 and was not retained. This branch is distinct from the outcome queries used in the final model.

3.2 Outcome queries and severity supervision

Four learnable outcome queries cross-attend the encoded tokens through one multi-head attention layer, adapting the label-query interface used by FORMED (Huang et al., 2026). Query and token inputs are layer-normalised, and the attention output is added to the corresponding original query. Each resulting vector feeds one head: binary risk, ordinal severity, continuous pH, or base deficit. Figure 2 shows how the heads share the encoder while retaining separate outcome representations.

The binary head produces a logit a and risk $\sigma(a)$. Its binary cross-entropy uses a positive-class weight equal to the training-set ratio of negative to positive records. The ordinal head instead predicts four nested events,

$$y_k = \mathbb{1}[\mathrm{pH} < \tau_k], \qquad (\tau_1, \tau_2, \tau_3, \tau_4) = (7.20, 7.15, 7.10, 7.05). \tag{1}$$

A shared score s and threshold-specific biases give $P_k = \sigma(s - b_k)$, following the shared-score ordinal formulation (Cao et al., 2020). For ordered biases $b_1 < b_2 < b_3 < b_4$, the predictions satisfy

$$P_1 \geq P_2 \geq P_3 \geq P_4. \tag{2}$$

The implementation initialises the biases to $(-2.0, -0.7, 0.7, 2.0)$ and learns them jointly. We assess their resulting prediction order empirically; the reported evaluations satisfy Equation 2 throughout. The ordinal loss is the mean binary cross-entropy over the four targets. Unlike the primary binary label, this **nested supervision** distinguishes multiple regions of the pH spectrum.

The continuous heads each predict a mean μ and log standard deviation ℓ, with ℓ clipped to $[-4, 2]$. Targets are standardised using training-fold statistics. For a standardised outcome u, the Gaussian negative log-likelihood is

$$\mathcal{L}_{\mathrm{G}}(u; \mu, \ell) = \ell + \tfrac{1}{2}(u - \mu)^2 e^{-2\ell} + \tfrac{1}{2}\log(2\pi). \tag{3}$$

Base-deficit loss is averaged only over records with an observed measurement; missing outcomes contribute no synthetic target. The complete severity objective is

$$\mathcal{L}_{\mathrm{out}} = \mathcal{L}_{\mathrm{bin}} + \mathcal{L}_{\mathrm{ord}} + \mathcal{L}_{\mathrm{pH}} + \mathcal{L}_{\mathrm{BDecf}}. \tag{4}$$

These objectives encourage a shared representation of severity, while the primary binary head remains the source of the reported acidaemia risk. The ordinal constraint relates the four ordinal outputs; it does not impose equality between the separate binary and ordinal heads.

3.3 Quality-conditioned dual-view distillation

QvD uses a **raw student view** and a **repaired teacher view**. Reliability is computed directly from the FHR trace, UC trace, and FHR validity mask. Seven components describe validity, missing-run duration, flatlines, jumps, physiological range, UC saturation, and local missingness. Missing-run reliability decays exponentially with a 30 s time constant. Flatlines use a 60-sample rolling standard deviation below 2, jumps use $|\Delta\mathrm{FHR}| > 25$, the FHR range is $[50, 200]$, and UC saturation is flagged at 100 or above. The final component is one minus the local missing fraction. Their mean is set to zero at missing FHR samples and averaged within each patch to obtain $q_p \in [0, 1]$.

The teacher's FHR view linearly interpolates bounded missing runs of at most 120 s between observed endpoints. Longer runs and boundary gaps remain missing; a five-point median filter is applied at detected jumps. UC is unchanged in both views. Both encoders receive reliability computed from the original observations. Thus, the quality conditions enter through the reliability features and the rules defining the repair; the scalar q_p does not weight the final consistency loss. Raw missing samples retain their zero sentinel before training-fold normalisation.

The teacher is an EMA copy of the student encoder, updated after each optimiser step with decay 0.99. Let z_p^{raw} and z_p^{rep} be corresponding FHR-token representations. The consistency objective is

$$\mathcal{L}_{\mathrm{cons}} = \frac{1}{Pd} \sum_{p=1}^{P} \left\| z_p^{\mathrm{raw}} - \mathrm{sg}(z_p^{\mathrm{rep}}) \right\|_2^2, \qquad P = 225, \tag{5}$$

where sg stops gradients and $d = 128$. Averaging over both tokens and features matches the implementation. **Uniform patch weighting** retains consistency supervision at unreliable regions,

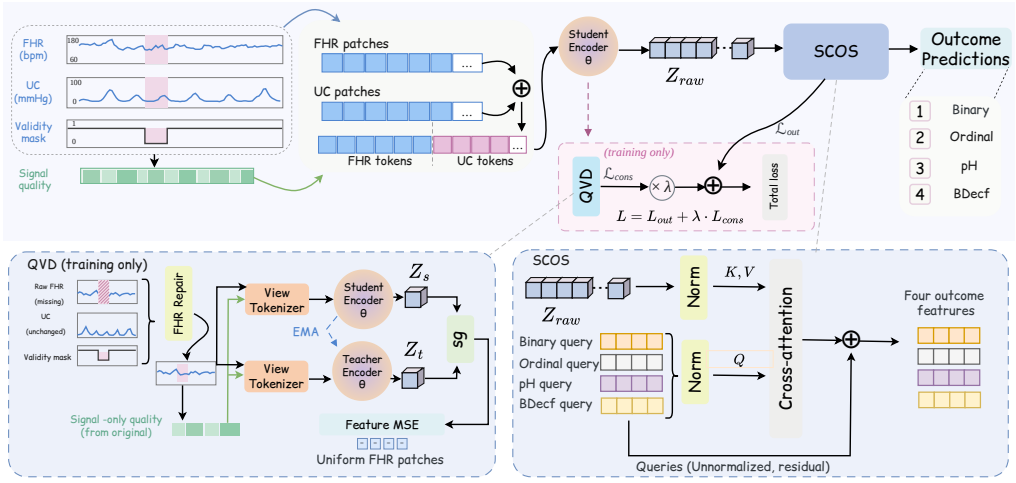

Figure 2: **FORGE-CTG overview.** *Top*: the FHR and UC channels are patchified and summed with a statistic embedding over (validity, mean, standard deviation, and the signal-quality score q_p); a small student encoder produces Z_{raw}, which SCOS decodes into four outcome views (binary, ordinal, pH, BDecf). *Bottom left*: QVD (training only): a conservatively repaired view is read by an EMA teacher while the student reads the raw view, and a feature MSE with uniform FHR-patch weight pulls Z_s toward $\text{sg}(Z_t)$; the total objective is $\mathcal{L} = \mathcal{L}_{\text{out}} + \lambda\mathcal{L}_{\text{cons}}$. *Bottom right*: SCOS in detail: four learnable outcome queries cross-attend the encoder output; the ordinal query compares a shared score against four threshold biases; the evaluated outputs satisfy the expected cumulative ordering.

where the two views can differ most. The reliability-weighted alternative is evaluated in the ablation study.

We optimise $\mathcal{L} = \mathcal{L}_{\text{out}} + 0.5\mathcal{L}_{\text{cons}}$ using AdamW with learning rate 10^{-3}, weight decay 10^{-4}, batch size 64, and gradient clipping at 5. Training runs for at most 50 epochs with validation-loss early stopping after 10 epochs without improvement. All student components are trained jointly from scratch. At inference, only the student and outcome heads are used; neither a repaired input nor an EMA teacher is required.

4 EXPERIMENTS

4.1 DATA AND EVALUATION PROTOCOL

CTU-UHB contains 552 intrapartum recordings (Chudáček et al., 2014); 551 are available in the local experimental subset because record 1500 lacks its signal file. The endpoint pH < 7.15 gives 105 positives (19.06%). Five stratified record-level folds keep all samples from a record together. Each outer training partition contains 440 or 441 records, with 20% held out for validation. Signal and continuous-target statistics are fitted on the remaining training records. Post-delivery variables, including pH, base deficit, Apgar scores, delivery mode, and birth weight, are excluded from model inputs.

The benchmark contains 19 configurations: three clinical-feature models (logistic regression, random forest, and XGBoost (Chen & Guestrin, 2016)); MiniRocket (Dempster et al., 2021); eight general or medical time-series models (Transformer, PatchTST, iTransformer, TimesNet, InceptionTime, Medformer, GRU, and LSTM (Hochreiter & Schmidhuber, 1997)); four missingness-oriented models (GRU-D, Raindrop, SAITS, and mTAN); two pretrained-model configurations (CTGDL-PatchTST and FORMED); and FORGE-CTG. Neural baselines use task-specific implementations or classification adaptations, detailed in Appendix D. Pretrained encoders are frozen in the evaluated configurations. Our main result averages five seeds per record; baseline results are single-seed. The comparison therefore measures the reported configurations under shared folds, rather than a seed-matched architecture ranking.

Table 1: CTU-UHB benchmark ($n = 551$, 105 positives). Intervals are record-bootstrap 95% CIs. Our row is a five-seed ensemble; other rows are single-seed configurations. Bold denotes the best value in each metric.

Model	AUROC ↑	95% CI	AUPRC ↑	Brier ↓
FORGE-CTG	**0.7283**	[0.673, 0.779]	**0.4226**	0.2127
SAITS	0.7190	[0.668, 0.767]	0.3520	0.2386
Clinical LR	0.7099	[0.651, 0.762]	0.4047	0.2032
mTAN	0.7064	[0.650, 0.763]	0.3812	0.2146
PatchTST	0.6900	[0.635, 0.742]	0.3400	0.2145
Random forest	0.6817	[0.625, 0.734]	0.3262	0.1590
XGBoost	0.6766	[0.619, 0.732]	0.3510	0.1597
Raindrop	0.6704	[0.613, 0.724]	0.3540	0.2301
Transformer	0.6703	[0.613, 0.727]	0.3488	0.2136
CTGDL-PatchTST	0.6683	[0.610, 0.728]	0.3530	0.2265
InceptionTime	0.6579	[0.601, 0.710]	0.2842	0.2767
MiniRocket	0.6478	[0.583, 0.705]	0.2885	0.2049
Medformer	0.5887	[0.523, 0.651]	0.2743	**0.1524**
iTransformer	0.5587	[0.496, 0.622]	0.2371	0.2441
GRU	0.5414	[0.481, 0.604]	0.2220	0.2645
TimesNet	0.5230	[0.464, 0.583]	0.1980	0.2539
LSTM	0.5177	[0.455, 0.579]	0.1989	0.2726
FORMED	0.5029	[0.442, 0.566]	0.1898	0.2510
GRU-D	0.4914	[0.431, 0.553]	0.2009	0.2735

CTGDL-PatchTST uses 550 records (104 positives); its paired comparison uses the common records. All names refer to the evaluated configurations described in Appendix D. Paired tests appear in Table 4.

AUROC is the primary metric and AUPRC measures precision–recall ranking. We also report Brier score and expected calibration error (ECE). Marginal 95% confidence intervals use 2,000 record-level bootstrap replicates. Paired comparisons resample the same records for both models and apply Holm–Bonferroni correction across 18 comparisons. Model selection uses validation loss. Appendix C distinguishes pooled ensemble results from averages across individual folds and seeds.

4.2 CTG BENCHMARK RESULTS

Table 1 shows that FORGE-CTG has the highest observed AUROC (0.7283, 95% CI 0.673–0.779) and AUPRC (0.4226) among the evaluated configurations. Its paired advantage is significant against eight alternatives after Holm correction. Differences against SAITS, clinical logistic regression, and mTAN are not significant ($p_{\text{Holm}} = 0.981$), so the observed rank does not establish superiority over these strong baselines.

Ranking and calibration differ: clinical tree models and Medformer have lower Brier scores than FORGE-CTG despite lower AUROC. We therefore use within-backbone experiments to isolate the effects of supervision and distillation.

4.3 COMPONENT AND DESIGN ABLATIONS

Table 2 compares configurations using the same backbone and three-seed evaluation. Binary supervision gives 0.7028 AUROC; adding SCOS increases this to 0.7200, and adding QVD gives 0.7342. The successive gains are 0.0172 and 0.0142.

Uniform weighting yields 0.7300 clean AUROC and a missingness slope of -0.0623, compared with 0.7217 and -0.1176 for reliability weighting. The degradation magnitude is $1.9\times$ smaller. Reliability informs the teacher view without attenuating the student–teacher loss at difficult patches. Interpolation has a flatter slope (-0.0010) but lower clean AUROC (0.7170), making uniform distillation the stronger clean–robustness balance.

Table 2: Mechanism ablations on CTU-UHB. Panels use their respective three-seed protocols; absolute values should be compared within a panel. Slopes are AUROC change per unit random-missing fraction.

(a) Controlled component ladder

Configuration	AUROC ↑	Change from preceding row
Binary supervision	0.7028	—
+ ScoS	0.7200	+0.0172
+ QvD (Forge-CTG)	**0.7342**	+0.0142

(b) Input-view and consistency variants

Configuration	Clean AUROC ↑	Missingness slope
Local interpolation	0.7170	−0.0010
Mask-only reconstruction	0.7040	−0.1625
Attention imputation	0.6932	−0.5231
Instance normalisation	0.7098	−0.1704
QvD, reliability weights	0.7217	−0.1176
QvD, uniform weights	**0.7300**	−0.0623

The attention-imputation control is a two-layer masked reconstruction model, distinct from the benchmark SAITS classification adaptation. The uniform variant has the highest clean AUROC; interpolation has the slope closest to zero.

5 Robustness, Label Efficiency, and Transfer

5.1 Robustness to missing observations

We corrupt only test records, masking FHR samples independently at 5, 10, 20, and 40% or inserting one gap of 15, 30, 60, or 120 s. Reliability features are recalculated with the original training-fold statistics, and the student still receives the raw view. The fitted model is unchanged, so the slope measures test-time sensitivity to missing observations.

With ScoS alone, AUROC changes at -0.3126 per unit increase in random missingness. Adding QvD changes the slope to -0.0678, a $4.6\times$ reduction in degradation magnitude under fixed outcome losses, folds, and corruptions. In the separate gap test, the full model remains between 0.731 and 0.737 AUROC from 15 to 120 s. Random masking spreads uncertainty over many patches, whereas a gap removes one interval. Similar stability under both patterns supports a **consistent robustness gain**.

The view ablation explains the effect. Uniform consistency raises clean AUROC from 0.7217 to 0.7300 and reduces slope magnitude from 0.1176 to 0.0623 relative to reliability weighting. Reliability informs the teacher view, but using it again as a loss weight weakens supervision where raw and repaired inputs differ most. Uniform weighting keeps those patches in the learning signal. Interpolation has the flattest missingness slope but lower clean AUROC, and its AUROC drops to 0.5017 and 0.4442 under the $0.5\times$ and $2\times$ amplitude tests. It removes missing samples directly but does not protect against distortions outside that repair rule. QvD gives the stronger balance across clean and corrupted inputs.

5.2 Label efficiency and subgroup variation

We subsample training and validation records and keep each test fold intact. With 50% of the development labels, the three-seed model obtains ***0.7291 AUROC*** (95% CI 0.6707–0.7814), versus 0.7268 (0.6704–0.7793) with all labels. This supports a useful **half-label operating point**, although it is not an equivalence test. AUROC falls to 0.6877 at 25% and 0.5984 at 10%, where the smallest folds contain only about six positive training records. The break below 50% is consistent with positive-case scarcity. The full-label three-seed estimate is within 0.0015 of the five-seed result, so seed aggregation is unlikely to explain the half-label point estimate.

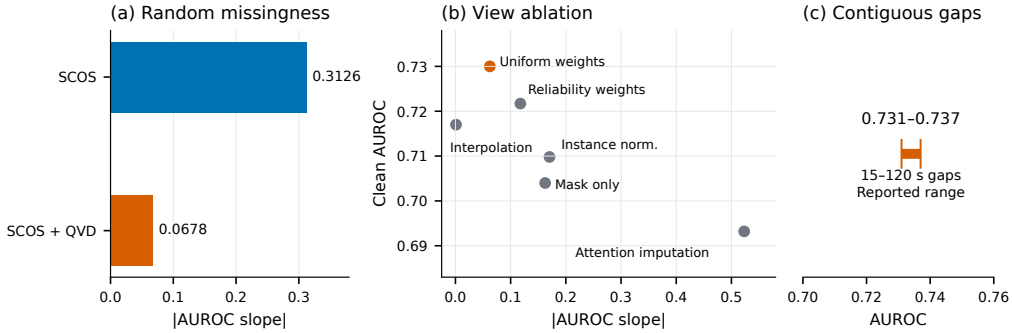

Figure 3: CTG robustness. (a) Random-missingness slope magnitude; smaller is better. (b) Clean AUROC and slope magnitude in the input-view ablation. (c) AUROC range across 15–120 s gaps.

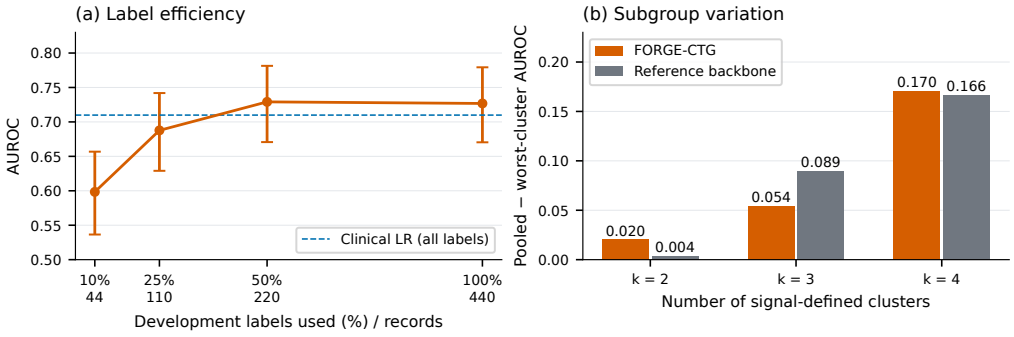

Figure 4: Label efficiency and signal-defined variation. (a) AUROC with record-bootstrap 95% intervals; counts denote development records in outer fold 0. (b) Pooled minus worst-cluster AUROC.

For $k = 2, 3, 4$ signal-defined clusters, the pooled-to-worst-cluster gaps are 0.0203, 0.0543, and 0.1704, versus 0.0035, 0.0893, and 0.1663 for the reference. The gap narrows at $k = 3$ but not at the other resolutions; at $k = 4$, the 36-record worst cluster reaches 0.5579 AUROC. This hard tail limits conclusions from the pooled score. These are signal clusters, not demographic groups.

5.3 Transfer to adult ECG

We test the same learning principles on PTB-XL (Wagner et al., 2020) ($n = 3000$) and CPSC2018 (Liu et al., 2018) ($n = 2826$). Each ECG contributes 20 patches from each of 12 leads. The adaptation retains the shared encoder and outcome queries, removes the base-deficit head, and uses ECG-specific reliability features. PTB-XL folds are patient-disjoint; the CPSC2018 subset has one record per patient. Diagnostic categories form the ordinal target, so this tests ordered supervision rather than a validated clinical severity scale.

SCOS reduces ordinal MAE from 1.237 to 0.943 on PTB-XL (***23.8%***) and from 1.191 to 0.942 on CPSC2018 (***20.9%***), with rank consistency of 1.0. Agreement across two category distributions supports transfer of the supervision design. Clean binary AUROC does not improve: the full model obtains 0.8879 versus 0.8952 on PTB-XL and 0.8471 versus 0.8507 on CPSC2018; the PTB-XL difference is significant ($p = 0.022$). The gain is **target-specific**: the outcome spectrum improves the constructed ordering, not binary discrimination.

At 8 s gaps, FORGE-CTG reaches 0.7617 mean AUROC on PTB-XL and 0.7044 on CPSC2018, leading the three framework configurations on both datasets. It also leads at 40% random missingness on CPSC2018. The 8 s margins over binary supervision are 0.0059 and 0.0031; over SCOS, they are 0.0010 and 0.0055. These differences are smaller than the run-level standard deviations of 0.0242 and 0.0210. We therefore read the shared direction as **stress-test consistency**, while the ordinal reduction remains the stronger transfer result.

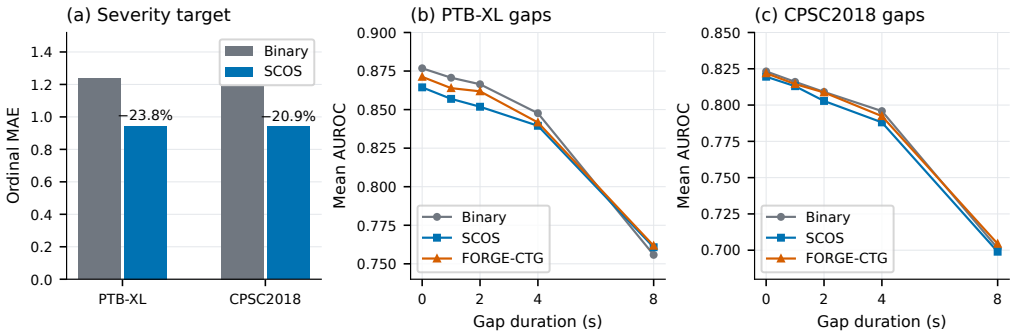

Figure 5: ECG transfer. (a) Ordinal MAE for constructed severity targets. (b,c) Mean AUROC under contiguous gaps; the full model leads the three configurations at 8 s.

6 Conclusion

FORGE-CTG combines **outcome-spectrum supervision** and **dual-view distillation** in a compact CTG encoder. Four outcome queries connect binary risk to nested thresholds and continuous metabolic outcomes, while a raw-input student learns from an EMA teacher with a conservatively repaired FHR view. The resulting model retains a simple inference path and uses signal-derived quality information without requiring post-delivery inputs.

On CTU-UHB, the model attains 0.7283 AUROC and 0.4226 AUPRC, and controlled ablations support contributions from both mechanisms. Adding QvD reduces the random-missingness degradation magnitude by $4.6\times$ relative to SCOS alone. The half-label experiment yields a point estimate close to the full-label result, while ECG experiments show that the outcome-spectrum principle can improve constructed ordinal targets across cardiac signals. These findings make supervision structure and training-view design useful complements to backbone development for physiological monitoring.

Taken together, the results suggest that the main opportunity is not simply to enlarge the encoder, but to align what the representation is asked to predict with what can be trusted in the input. The outcome queries encourage a shared latent state to retain ordered and continuous metabolic information, whereas the paired views specify which perturbations should leave that state stable. Their combination therefore addresses two distinct failure modes—label compression and unreliable observation—without adding post-processing to the inference pipeline. The modest margins over strong baselines and the variability across stress tests nevertheless argue for viewing these mechanisms as complementary inductive biases rather than as a substitute for broader data or external validation.

Limitations. CTG evaluation is restricted to one outcome-labelled database and is retrospective. The main ensemble is not seed-matched to the baselines, its advantage over the strongest alternatives is not significant, and calibration and worst-cluster performance remain limited. ECG severity labels are constructed, and the ordinal improvement does not imply better binary diagnosis. External outcome-labelled CTG evaluation, calibration, and prospective assessment are the next steps toward establishing clinical usefulness.

Future work should evaluate discrimination, severity ordering, calibration, and robustness together across acquisition settings and outcome prevalences. In particular, external cohorts could test whether reliability features remain informative under different monitors and clinical workflows, while prospective studies could examine whether severity-aware predictions improve decision support rather than only retrospective ranking. Extending the framework to uncertainty-aware calibration and longer patterns of structured signal loss would further clarify when the teacher view provides useful guidance. These steps are needed to determine whether the observed gains extend beyond a retrospective benchmark.

AI USE STATEMENT

Generative AI tools assisted with the initial identification and survey of relevant literature, code for data preprocessing, and language editing and formatting of the manuscript. The authors manually reviewed the literature and checked the preprocessing code and its outputs. Research questions, hypotheses, study design, the model methodology, experiments, and scientific conclusions were developed and assessed by the authors. AI was not used to generate synthetic experimental data or results. The authors take responsibility for the final text, citations, code, analyses, and claims, including all AI-assisted material.

ETHICS STATEMENT

This study uses public, de-identified CTU-UHB, PTB-XL, and CPSC2018 records. No new patient data were collected and no individuals were contacted. The analysis is retrospective and evaluates risk ranking and outcome prediction; it does not establish a clinical intervention policy. Calibration and signal-defined subgroup results are reported alongside discrimination because pooled AUROC alone is insufficient to assess clinical suitability. Further external and prospective evaluation is needed before these predictions can inform patient care.

REPRODUCIBILITY STATEMENT

Section 3 specifies the architecture, outcome targets, repair rules, and training objective. Section 4 describes the record-level splits and statistical comparisons, and Appendix C provides preprocessing, hyperparameters, seed counts, aggregation conventions, and implementation details. Appendix D identifies the evaluated baseline adaptations. Additional benchmark, paired-comparison, and transfer results appear in Appendices A–B. These descriptions distinguish the main five-seed ensemble from three-seed mechanism studies and from mean-per-run ECG analyses.

REFERENCES

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. doi: 10.1016/j.patrec.2020.11.008.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Václav Chudáček, Jiří Spilka, Miroslav Burša, Petr Jankú, Lukáš Hruban, Michal Huptych, and Lenka Lhotská. Open access intrapartum CTG database. *BMC Pregnancy and Childbirth*, 14:16, 2014. doi: 10.1186/1471-2393-14-16.

Angus Dempster, Daniel F. Schmidt, and Geoffrey I. Webb. MiniRocket: A very fast (almost) deterministic transform for time series classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 248–257, 2021. doi: 10.1145/3447548.3467231.

Wenjie Du, David Côté, and Yan Liu. SAITS: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023. doi: 10.1016/j.eswa.2023.119619.

Naomi Fridman and Berta Ben Shachar. A foundation model approach for fetal stress prediction during labor from cardiotocography (CTG) recordings. *arXiv preprint arXiv:2601.06149*, 2026. URL https://arxiv.org/abs/2601.06149.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Nan Huang, Haishuai Wang, Zihuai He, Marinka Zitnik, and Xiang Zhang. Repurposing foundation model for generalizable medical time series classification. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8707924df5e207fa496f729f49069446-Abstract-Conference.html.

Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and François Petitjean. InceptionTime: Finding AlexNet for time series classification. *Data Mining and Knowledge Discovery*, 34(6):1936–1962, 2020. doi: 10.1007/s10618-020-00710-y.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=cGDAkQo1C0p.

Baijiong Lin and Yu Zhang. LibMTL: A python library for deep multi-task learning. *Journal of Machine Learning Research*, 24(209):1–7, 2023. URL https://www.jmlr.org/papers/v24/22-0347.html.

Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, Jianqing Li, and Eddie Ng Yin Kwee. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 8(7):1368–1373, 2018. doi: 10.1166/jmihi.2018.2442.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06625.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.14730.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=4c0J6lwQ4_.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/2017/hash/68053af2923e00204c3ca7c6a3150cf7-Abstract.html.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7:154, 2020. doi: 10.1038/s41597-020-0495-6.

Yihe Wang, Nan Huang, Taida Li, Yujun Yan, and Xiang Zhang. Medformer: A multi-granularity patching transformer for medical time-series classification. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/3fe2a777282299ecb4f9e7ebb531f0ab-Abstract-Conference.html.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02186.

Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.05357.

A DETAILED CTG RESULTS

Table 3 complements the main ranking with calibration and thresholded classification metrics. AUROC and AUPRC describe ranking; Brier score and ECE describe probability calibration. F_1 uses a Youden threshold selected from the pooled out-of-fold predictions and is a descriptive, retrospectively selected operating point. Specificity is reported at the empirical point reaching 90% sensitivity.

Table 3: Additional CTU-UHB metrics. ECE uses ten equal-width bins. Spec90 is specificity at 90% sensitivity; F_1 uses the pooled Youden threshold. Bold marks the best value per metric.

Model	Records	Brier ↓	ECE ↓	Spec90 ↑	F_1 ↑
FORGE-CTG	551	0.2127	0.2606	0.2780	0.4535
SAITS	551	0.2386	0.3055	**0.3543**	**0.4610**
Clinical LR	551	0.2032	0.2411	0.2825	0.4471
mTAN	551	0.2146	0.2548	0.2309	0.4365
PatchTST	551	0.2145	0.2596	0.2982	0.4032
Random forest	551	0.1590	0.0958	0.2534	0.4178
XGBoost	551	0.1597	0.1057	0.2713	0.4061
Raindrop	551	0.2301	0.2819	0.2601	0.4049
Transformer	551	0.2136	0.2374	0.2085	0.4214
CTGDL-PatchTST	550	0.2265	0.2835	0.2825	0.4041
InceptionTime	551	0.2767	0.3440	0.3027	0.3920
MiniRocket	551	0.2049	0.1671	0.1794	0.4073
Medformer	551	**0.1524**	**0.0279**	0.1637	0.3538
iTransformer	551	0.2441	0.2631	0.1233	0.3229
GRU	551	0.2645	0.3314	0.0874	0.2743
TimesNet	551	0.2539	0.3156	0.1233	0.3279
LSTM	551	0.2726	0.3373	0.1278	0.3281
FORMED	551	0.2510	0.3106	0.0830	0.2721
GRU-D	551	0.2735	0.3377	0.0695	0.1250

Table 4: Paired record-bootstrap comparison, with ΔAUROC defined as FORGE-CTG minus the comparator. Differences are bootstrap means. Holm correction covers all 18 comparisons.

Model	n_{common}	ΔAUROC	95% CI	p_{boot}	p_{Holm}
GRU-D	551	+0.2348	[0.1541, 0.3164]	<0.001	<0.001
FORMED	551	+0.2236	[0.1501, 0.3014]	<0.001	<0.001
LSTM	551	+0.2094	[0.1360, 0.2866]	<0.001	<0.001
TimesNet	551	+0.2031	[0.1244, 0.2817]	<0.001	<0.001
GRU	551	+0.1856	[0.1059, 0.2601]	<0.001	<0.001
iTransformer	551	+0.1685	[0.0930, 0.2444]	<0.001	<0.001
Medformer	551	+0.1388	[0.0791, 0.2008]	<0.001	<0.001
MiniRocket	551	+0.0803	[0.0253, 0.1343]	0.003	0.033
InceptionTime	551	+0.0702	[0.0136, 0.1245]	0.015	0.150
CTGDL-PatchTST	550	+0.0613	[0.0014, 0.1261]	0.044	0.312
Vanilla transformer	551	+0.0569	[−0.0005, 0.1117]	0.054	0.312
Raindrop	551	+0.0567	[0.0030, 0.1085]	0.039	0.312
XGBoost	551	+0.0516	[0.0014, 0.1019]	0.045	0.312
Random forest	551	+0.0462	[0.0060, 0.0861]	0.026	0.234
PatchTST	551	+0.0375	[−0.0091, 0.0835]	0.110	0.440
mTAN	551	+0.0206	[−0.0303, 0.0716]	0.424	0.981
Clinical LR	551	+0.0183	[−0.0174, 0.0559]	0.327	0.981
SAITS	551	+0.0085	[−0.0378, 0.0527]	0.706	0.981

B ECG TRANSFER RESULTS

PTB-XL and CPSC2018 provide separate tests of the outcome-query and dual-view principles. Table 5 reports pooled ensemble ranking, Table 6 reports paired differences, and Table 7 reports means over individual runs. The distinction matters because averaging AUROCs is different from calculating an AUROC after averaging predictions. All comparisons retain the single-seed baseline and three-seed framework convention of the transfer experiments.

Table 5: ECG pooled out-of-fold AUROC with record-bootstrap 95% CIs. Framework configurations average three seeds per record; other models use one seed.

	PTB-XL ($n = 3000$)		CPSC2018 ($n = 2826$)	
Model	AUROC	95% CI	AUROC	95% CI
Binary	0.8952	[0.884, 0.906]	**0.8507**	[0.837, 0.865]
SCOS	0.8833	[0.871, 0.894]	0.8497	[0.836, 0.863]
FORGE-CTG	0.8879	[0.876, 0.899]	0.8471	[0.833, 0.861]
GRU-D	**0.9041**	[0.893, 0.915]	0.8421	[0.828, 0.856]
SAITS	0.8948	[0.884, 0.906]	0.8392	[0.825, 0.853]
mTAN	0.8744	[0.862, 0.887]	0.8074	[0.792, 0.822]
PatchTST	0.8716	[0.859, 0.884]	0.8434	[0.829, 0.857]
Transformer	0.8604	[0.847, 0.873]	0.8272	[0.812, 0.842]
Raindrop	0.8587	[0.846, 0.872]	0.7681	[0.751, 0.784]
GRU	0.8265	[0.812, 0.841]	0.8478	[0.834, 0.862]
XGBoost	0.8191	[0.804, 0.834]	0.8299	[0.815, 0.845]
Clinical LR	0.7990	[0.782, 0.815]	0.7967	[0.780, 0.812]

Table 6: Paired ECG changes in pooled AUROC. Δ is the destination configuration minus the starting configuration. Point differences are bootstrap medians; intervals use the same direction; p is the paired-bootstrap value.

Dataset	Transition	ΔAUROC	95% CI	p
PTB-XL	Binary $\rightarrow$ SCOS	-0.0119	$[-0.0185, -0.0056]$	< 0.001
PTB-XL	SCOS $\rightarrow$ FORGE-CTG	$+0.0048$	$[-0.0009, +0.0111]$	0.102
PTB-XL	Binary $\rightarrow$ FORGE-CTG	-0.0072	$[-0.0133, -0.0007]$	0.022
CPSC2018	Binary $\rightarrow$ SCOS	-0.0011	$[-0.0097, +0.0077]$	0.799
CPSC2018	SCOS $\rightarrow$ FORGE-CTG	-0.0028	$[-0.0111, +0.0055]$	0.521
CPSC2018	Binary $\rightarrow$ FORGE-CTG	-0.0038	$[-0.0124, +0.0050]$	0.411

Table 7: ECG metrics averaged across five folds and three seeds. These are per-run means, distinct from the pooled ensemble AUROCs in Table 5. The final two rows use a slope against the index of the nine ordered corruption conditions, not a physical missingness fraction.

(a) Ordinal MAE and rank consistency

Dataset	Config	sev. MAE	mono	binary AUROC
PTB-XL	binary	1.237	1.000	0.8768
PTB-XL	SCOS	**0.943**	1.000	0.8644
PTB-XL	FORGE-CTG	0.952	1.000	0.8713
CPSC2018	binary	1.191	1.000	0.8232
CPSC2018	SCOS	**0.942**	1.000	0.8195
CPSC2018	FORGE-CTG	0.950	1.000	0.8219

(b) AUROC under representative corruption conditions

Dataset	Condition	binary	SCOS	FORGE-CTG
PTB-XL	clean	**0.8768**	0.8644	0.8713
PTB-XL	random missing 40%	**0.8514**	0.8442	0.8470
PTB-XL	gap 8 s	0.7558	0.7607	**0.7617**
CPSC2018	clean	**0.8232**	0.8195	0.8219
CPSC2018	random missing 40%	0.7898	0.7868	**0.7925**
CPSC2018	gap 8 s	0.7013	0.6990	**0.7044**
PTB-XL	Mean degradation slope	−0.00957	−0.00830	−0.00881
CPSC2018	Mean degradation slope	−0.00964	−0.00967	−0.00931

Historical comparison. An earlier experiment report records a temporal convolutional network at 0.9232 AUROC on PTB-XL and 0.9002 on CPSC2018. Its prediction files are unavailable for the paired analysis in Table 6. We accordingly restrict the transfer claim to ordinal severity supervision and the reported stress conditions, rather than overall ECG discrimination.

C Protocol and Reproducibility Details

C.1 CTG preparation and model configuration

FHR and UC are sampled at 4 Hz and cropped to the final 30 minutes. FHR zeros define missingness; UC zeros remain valid. FHR statistics use positive training samples and UC statistics use all training samples. Missing FHR values retain their sentinel before normalisation, and patch statistics respect the validity mask. Repair changes the repaired-view mask but not the reliability description of the original trace.

Each outer training partition reserves 20% for validation with the fold-index seed. The main result uses seeds 0–4 (25 runs); ablation, corruption, and reduced-label studies use seeds 0–2. Reduced-label fractions are applied separately to training and validation records.

Ordinal implementation. The ordinal score uses four trainable, increasingly initialised biases without sorting or reparameterisation during optimisation. Rank consistency is therefore checked empirically on adjacent cumulative predictions with tolerance 10^{-6}.

Training environment and aggregation. Experiments use PyTorch 2.3.1 with CUDA 12.1. Main CTG metrics average five seed-specific risks per record. Marginal bootstrap intervals resample records with replacement; paired intervals reuse the same indices for both methods. Holm correction is applied within each comparison family.

Table 8: Configuration of the final CTG model.

Component	Setting
Input / sampling	FHR and UC; 7,200 samples; 4 Hz
Patch length / stride	32 / 32 samples
Encoder	3 pre-norm layers; width 128; 4 heads
Feed-forward width / dropout	512 / 0.1
Decoder	4 outcome queries; one cross-attention layer
Optimiser	AdamW; learning rate 10^{-3}; weight decay 10^{-4}
Batch size / gradient clipping	64 / 5.0
Maximum epochs / patience	50 / 10 (validation loss)
Distillation / EMA decay	$\lambda = 0.5$ / 0.99
Repair duration / reliability decay	At most 120 s / 30 s
Reliability window	60 samples (15 s)
Gaussian log-scale interval	$[-4, 2]$
Main ensemble / ablations	5 seeds / 3 seeds, each over 5 folds

C.2 ECG PREPROCESSING AND TARGETS

All 12 ECG leads are pooled from 500 to 100 Hz; the first 10 s form 20 fifty-sample patches per lead. PTB-XL uses five patient-disjoint folds derived from official folds, CPSC2018 uses record-level stratification, and subset seed 2026 is fixed. Binary targets separate normal and abnormal records; the auxiliary target is an ordinal rank, with no base-deficit head.

PTB-XL ranks NORM, STTC, CD, HYP, and MI from 0 to 4. CPSC2018 ranks Normal=0; PAC/PVC=1; I-AVB/LBBB=2; STD/AF=3; and STE=4. This is an experimental ordering, not a validated severity scale. ECG repair covers gaps of at most 1 s and uses ECG-specific lead-off and amplitude checks. Transfer configurations use three seeds.

D EVALUATED BASELINE CONFIGURATIONS

Model names refer to the evaluated adaptations. General time-series models use record-level classification heads; CTGDL-PatchTST and FORMED use frozen pretrained encoders, with a trained outcome-query adapter for FORMED (Huang et al., 2026). In the supplied GRU-D forward pass, an all-observed mask leaves input imputation inactive. Raindrop and mTAN pool time to 128 positions; mTAN adds radial-basis time features. SAITS uses patch tokens and a classification objective. The two-layer attention-imputation control in Table 2(b) is instead trained for masked reconstruction, and the instance-normalisation control uses channel-wise instance statistics. Results therefore describe these configurations rather than unrestricted reproductions of their source methods.