Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

One-pass restoration and translation of cardiovascular signals with diffusion-trained state-space networks

Ugo Lomoio
University of catanzaro
Pietro Lio'
University of Cambridge
Pierangelo Veltri
University of Calabria
Pietro Hiram Guzzi
hguzzi@unicz.it

University of Catanzaro https://orcid.org/0000-0001-5542-2997

Research Article

Keywords: cardiovascular signals, diffusion models, state-space models, Mamba, signal restoration, cross-modal translation, real-time monitoring

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11731604/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

One-pass restoration and translation of cardiovascular signals with diffusion-trained state-space networks

Ugo Lomoio[1,2], Pietro Liò[3], Pierangelo Veltri[1,*], Pietro Hiram Guzzi[2]

[1]Department of Informatics Modelling and Systems, University of Calabria, Rende, Italy.
[2]Department of Surgical and Medical Sciences, Magna Graecia University of Catanzaro, Catanzaro, Italy.
[3]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK.

Abstract

Cardiovascular monitoring often relies on recordings degraded by noise or interrupted by signal loss, and some modalities may be unavailable. Existing learning-based approaches can recover missing waveforms, but unified diffusion models can require repeated network evaluations at inference. A bidirectional state-space network trained on diffusion-perturbed observations can instead reconstruct a waveform in one deterministic forward pass. DiBiMa-Cardio and its multimodal extension apply this approach to electrocardiography, photoplethysmography and arterial blood pressure across 57 configurations spanning denoising, imputation, forecasting, super-resolution and cross-modal translation. In a matched training comparison, diffusion perturbation and its step embedding reduced denoising error but increased super-resolution error in most configurations. In a separate inference comparison, one-pass reconstruction yielded lower normalised mean absolute error than the tested iterative sampler in 15 of 18 single-input configurations, with no significant difference in the other three. On 19,996 test windows, the multimodal model reduced normalised error relative to the base model in 53 of 57 configurations under the harder corruption protocol. The multimodal model processed a four-second window in approximately 105 milliseconds at batch size one on the test GPU. Restoring corrupted ECG

recovered much of the classification performance lost to corruption, while agreement between translated and recorded waveforms did not consistently preserve predicted diagnostic classes.

Keywords: cardiovascular signals, diffusion models, state-space models, Mamba, signal restoration, cross-modal translation, real-time monitoring

Introduction

Electrocardiography (ECG), photoplethysmography (PPG) and arterial blood pressure (BP) capture complementary aspects of cardiovascular function: electrical activation, peripheral pulse dynamics and haemodynamic load [1, 2]. Motion, poor sensor contact and low perfusion can degrade recorded waveforms, while interruptions leave gaps. Invasive arterial pressure monitoring may also be impractical, leaving the pressure waveform unavailable [3–6]. These constraints are particularly relevant outside hospitals, for example when older adults are monitored remotely with wearable devices [7, 8]. Reconstruction methods must therefore adapt to the signals that remain available and process each acquired window quickly enough for repeated use (Fig. 1a–c).

Learning-based methods can use one cardiovascular signal to reconstruct another. CardioGAN uses an adversarial generator with time- and frequency-domain discriminators to synthesise ECG from PPG [9], while the Region-Disentangled Diffusion Model (RDDM) uses region-specific perturbation and denoising for the same translation task [10]. HIPNO uses symmetry-aware physics-informed neural operators to infer arterial pressure and vascular mechanics from non-invasive ECG and PPG [11]. UniCardio extends the scope of reconstruction through a unified diffusion transformer that handles denoising, imputation and translation across ECG, PPG and BP under different combinations of observed signals [12]. This broader coverage makes UniCardio a reference for testing how reconstruction depends on the available modalities and the inference procedure.

Work on time-series reconstruction offers a complementary perspective. Conditional Score-based Diffusion for Imputation (CSDI) uses observed values to condition probabilistic imputation [13], and SSSD combines conditional diffusion with structured state-space layers for imputation and forecasting [14]. In biomedical signal generation, SSSD-ECG synthesises 12-lead ECG conditioned on diagnostic statements, and BioDiffusion supports unconditional, label-conditioned and signal-conditioned synthesis [15, 16]. Iterative diffusion sampling incurs the cost of repeated network evaluations. DDIM reduces the number of sampling steps, and consistency models enable one-step generation through a dedicated training formulation [17, 18]. For cardiovascular reconstruction, a practical question is whether a model trained with diffusion perturbations can recover the target directly from the available observations, and how the answer varies with the task and severity of corruption.

Consider a four-second window in which the ECG briefly drops out while PPG remains available. The observed ECG samples on either side of the gap and the concurrent PPG provide complementary evidence about the missing segment (Fig. 1c,d). An observation mask identifies the samples to recover. Selective state-space models process sequences in an input-dependent manner [19, 20], and bidirectional scans can use the ECG context on both sides of the gap within the acquired window [21].

DiBiMa-Cardio implements this approach with a bidirectional Mamba-2 encoder–decoder. During training, diffusion perturbations are applied to the observed inputs and the network learns to reconstruct the clean target. At inference, the network maps the available signals to a reconstructed waveform in one deterministic forward pass. The multimodal extension, DiBiMa-Cardio-M, adds fusion modules for clean auxiliary signals, such as the PPG in the ECG gap example (Fig. 2a,b). The framework covers 57 configurations spanning imputation, forecasting, denoising, super-resolution and cross-modal translation.

Two controlled comparisons examine the role of diffusion. One tests how perturbing inputs during training affects reconstruction, and the other compares one-pass inference with iterative sampling using the same trained model. Benchmark comparisons with UniCardio examine how accuracy and computational cost vary with corruption and signal availability. Downstream experiments on PTB-XL and MIMIC PERform AF, together with an analysis of predicted diagnostic classes before and after cross-modal translation, assess how waveform fidelity relates to information useful for classification [22, 23].

Results

A unified framework for one-pass cardiovascular reconstruction

DiBiMa-Cardio combines modality-specific bidirectional Mamba-2 encoders, a cross-modal attention block and routed decoders to reconstruct cardiovascular waveforms in one deterministic forward pass (Fig. 2a). Observation masks identify available samples, while a template prior derived from the observed target waveform provides additional conditioning for imputation and forecasting. The base model contains 3.87 million parameters. DiBiMa-Cardio-M extends this architecture with multimodal fusion modules, increasing the parameter count to 5.36 million and enabling fine-tuning with clean auxiliary signals alongside a corrupted target (Fig. 2b).

Evaluation covered 57 configurations: imputation, forecasting, denoising and super-resolution for each of three target modalities, with zero, one or two auxiliary signals, together with six single-source and three dual-source translations (Fig. 2c). UniCardio-M, the released multimodal UniCardio checkpoint evaluated with six-step DDIM sampling and averaging over six draws, supported 45 configurations; the remaining 12 involved super-resolution.

Comparisons used the 256-window UniCardio benchmark subset under two corruption protocols (Fig. 2d). The standard protocol retained the benchmark masks and used the training-level noise specified in Methods. The hard protocol increased imputation gaps and forecast horizons to 2.0 s and doubled the noise amplitude. The

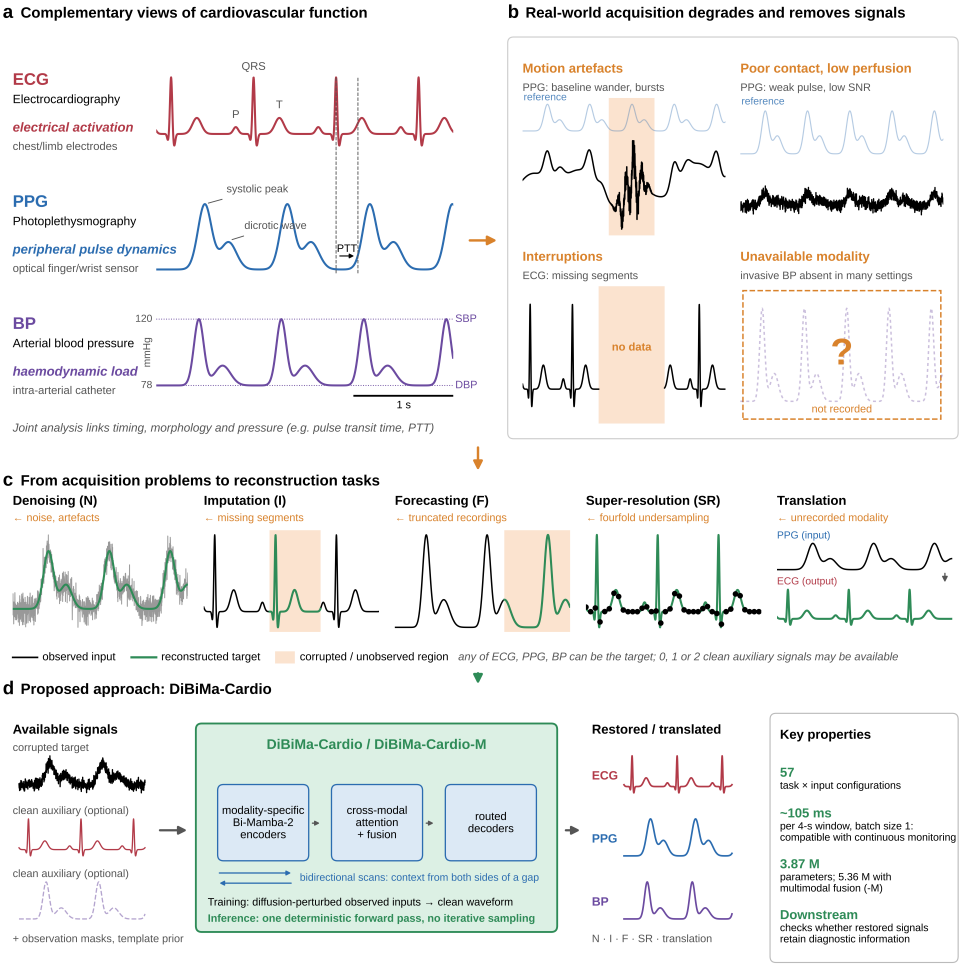


Fig. 1 Cardiovascular signal reconstruction under real-world monitoring conditions. Detailed panel descriptions are provided in Figure legends.

longer gaps and horizons were within DiBiMa-Cardio-M's fine-tuning range but outside UniCardio-M's training range; the doubled noise exceeded the training level of all models. Results are reported separately for the two protocols and complemented by corruption-severity sweeps.

Diffusion training and one-pass inference have task-dependent effects

A matched, smaller-scale comparison between BiMa and DiBiMa-Cardio examined the contribution of diffusion perturbation and the associated step embedding while retaining the same architecture and training data (Fig. 3a). Evaluation used 226 benchmark

windows outside these models' training and validation records. Under the hard protocol, diffusion training reduced normalised mean absolute error (nMAE) in 10 of 12 denoising configurations, with no significant difference in the remaining two and a median error ratio of 0.57. Super-resolution showed the opposite pattern: diffusion training increased nMAE in eight configurations, with no significant difference in four and a median ratio of 1.75. Across all 57 configurations, diffusion training produced 19 significant improvements, 17 deteriorations and 21 comparisons without a significant difference. Its clearest benefit was therefore in denoising.

To isolate the effect of the inference procedure, the same trained DiBiMa-Cardio model was run in two ways (Fig. 3b). The direct mode reconstructed each window in a single deterministic forward pass. The iterative mode used six DDIM steps per draw and averaged the reconstructions from six draws. Under the standard protocol, the direct mode yielded significantly lower nMAE in 15 of 18 single-input configurations, with no significant difference in the other three. Its PCC was significantly higher in 12 configurations and lower in one, with no significant difference in five. At batch size eight, median processing time in these evaluation runs was 16.8 ms per window for one-pass inference and 534 ms for iterative sampling.

Multimodal fine-tuning was examined through the cumulative trajectory leading to DiBiMa-Cardio-M (Fig. 3c). Fine-tuning without additional modules changed the median nMAE ratio relative to UniCardio-M from 1.23 to 1.22 across the 45 shared hard-protocol configurations. The ratios associated with the gated bottleneck, early fusion, decoder fusion, modality- and state-aware attention, and final template-prior dropout stages were 1.15, 1.09, 1.00, 0.96 and 0.95, respectively. Because the stages used different training budgets and lacked equal-step controls, these values describe the model-development trajectory rather than isolate individual module contributions. The final model had significantly lower nMAE than DiBiMa-Cardio in 53 of 57 hard-protocol configurations, with a median ratio of 0.62, and in all 57 standard-protocol configurations (Fig. 3d). The four hard-protocol exceptions involved ECG denoising.

Comparative performance depends on signal availability and corruption

Under the hard protocol, DiBiMa-Cardio-M achieved significantly lower nMAE than UniCardio-M in 22 of the 45 shared configurations, higher nMAE in nine, and no significant difference in 14 (record-clustered paired tests with Benjamini–Hochberg correction, $q < 0.05$; Fig. 4a,c). The median per-configuration nMAE ratio was 0.95. PCC favoured DiBiMa-Cardio-M in 29 configurations and UniCardio-M in six, with no significant difference in 10 (Fig. 4b,c).

The largest advantages occurred when only the corrupted target was available. DiBiMa-Cardio-M achieved significantly lower nMAE in all nine single-input restoration configurations, with a median ratio of 0.38 relative to UniCardio-M. Imputation nMAE was 0.122 versus 0.304 for PPG, 0.097 versus 0.199 for ECG, and 0.120 versus 0.279 for BP. Denoising nMAE was 0.091 versus 0.515, 0.083 versus 0.216, and 0.088 versus 0.416, respectively. PPG forecasting yielded nMAE of 0.131 versus 0.386.

The base DiBiMa-Cardio model also outperformed UniCardio-M in all nine configurations, with a median ratio of 0.62, indicating that this advantage was present before multimodal fine-tuning.

Clean auxiliary signals changed the comparative pattern. DiBiMa-Cardio-M achieved significantly lower nMAE in eight of nine auxiliary-conditioned denoising configurations, with median ratios of 0.20 for one auxiliary signal and 0.22 for two. For imputation and forecasting with two auxiliary signals, UniCardio-M achieved significantly lower nMAE in two of three configurations for each task, with no significant difference in the remaining configuration. Median DiBiMa-Cardio-M/UniCardio-M ratios were 1.22 for imputation and 1.17 for forecasting.

For single-source translation, DiBiMa-Cardio-M achieved lower nMAE in three of six configurations and no significant difference in the other three, with a median ratio of 0.94. PCC was significantly higher in all six. Dual-source translations showed no significant nMAE difference, with a median ratio of 0.99, but significantly higher PCC for DiBiMa-Cardio-M in all three configurations. In the 12 super-resolution configurations unsupported by UniCardio-M, DiBiMa-Cardio-M consistently improved on DiBiMa-Cardio; single-input ECG super-resolution yielded nMAE of 0.043 versus 0.085.

Under the standard protocol, DiBiMa-Cardio-M achieved lower nMAE than UniCardio-M in 14 configurations and higher nMAE in 16, with no significant difference in 15 and a median ratio of 1.00 (Supplementary Figs. 4 and 5; Supplementary Table 5). Its single-input advantage persisted in eight of nine configurations, whereas UniCardio-M achieved lower nMAE in 15 of 27 auxiliary-conditioned restoration configurations. The hard-protocol advantage was thus concentrated in particular operating regimes.

Corruption sweeps reveal distinct performance profiles

Sweeps over noise amplitude, gap length and forecast horizon examined performance across corruption levels (Fig. 4d). For single-input imputation and forecasting, DiBiMa-Cardio-M had the lowest mean nMAE at every evaluated gap and horizon length from 0.5 to 2.0 s across all six target–task combinations. UniCardio-M's single-input forecasting error varied non-monotonically, with mean nMAE of 0.37 at 0.5 s, 0.21 at 1.0 s and 0.32 at 2.0 s, compared with 0.075–0.119 for DiBiMa-Cardio-M over the evaluated range.

With clean auxiliary signals, UniCardio-M had the lowest mean imputation error for gaps from 1.0 to 2.0 s: nMAE ranged from 0.039 to 0.052, compared with 0.049–0.058 for DiBiMa-Cardio-M. DiBiMa-Cardio-M had the lowest error at 0.5 s. Both models showed relatively small changes in auxiliary-conditioned imputation error as gaps lengthened.

At noise scales $\alpha = 1$, 2 and 3, where $\alpha = 1$ denotes the DiBiMa-Cardio training noise level, mean single-input nMAE was 0.045, 0.087 and 0.200 for DiBiMa-Cardio-M, compared with 0.094, 0.386 and 0.550 for UniCardio-M. At $\alpha = 0$, UniCardio-M's denoising route produced nMAE of approximately 0.65. This clean-input condition was outside its denoising training distribution, whereas DiBiMa-Cardio training included

clean inputs. The sweeps therefore describe performance across differing training exposures as well as corruption severities.

One-pass reconstruction reduces processing cost

Windows selected by proximity to median error ranks illustrate the quantitative findings (Fig. 5a). In the displayed single-input imputation and forecasting examples, DiBiMa-Cardio-M more closely reproduced beat timing and morphology, while the denoising example showed greater suppression of added noise. Auxiliary-conditioned examples showed smaller differences between DiBiMa-Cardio-M and UniCardio-M.

Dedicated measurements on an NVIDIA RTX 4000 Ada GPU showed substantial throughput gains (Fig. 5b–d). At batch size 32, amortised processing time per 4-s window was 3.2 ms for DiBiMa-Cardio-M and 4.5 ms for DiBiMa-Cardio, compared with 324 ms for UniCardio-M using six-step DDIM and six draws. At batch size eight, the corresponding times were 13.0, 11.7 and 321 ms. At batch size one, they were 105, 94 and 472 ms. A single-draw UniCardio-M configuration required approximately 80 ms, although its reconstruction accuracy was not evaluated. The reported gains therefore apply to the sampling configurations used in the accuracy comparisons. Across the four timed settings, the maximum batch-one processing time for DiBiMa-Cardio-M was 139 ms per 4-s window. These measurements quantify processing after window acquisition.

DiBiMa-Cardio-M had the smallest measured GPU memory footprint: 92 MiB at batch size eight and 275 MiB at batch size 32, compared with approximately 0.39 and 1.4 GiB for UniCardio-M, and 2.2 and 8.7 GiB for DiBiMa-Cardio. Across the 45 shared hard-protocol configurations, mean nMAE was 0.088 for DiBiMa-Cardio-M, 0.151 for DiBiMa-Cardio and 0.183 for UniCardio-M. DiBiMa-Cardio-M thus combined lower aggregate error with reduced processing cost in the evaluated configurations (Supplementary Table 31).

Full-test evaluation confirms improvements over the base model

Evaluation was extended to 19,996 test windows from 7,264 records for DiBiMa-Cardio and DiBiMa-Cardio-M (Supplementary Tables 16–29). UniCardio-M was not evaluated on this full set. The window-level split permits recordings to contribute windows to different partitions, so this analysis does not establish record- or patient-independent generalisation.

Per-configuration nMAE on the benchmark subset and full test set had Spearman correlations of at least 0.98 for both models under both protocols. Hard-protocol subset errors were nevertheless lower at the median by 4% for DiBiMa-Cardio and 10% for DiBiMa-Cardio-M. These differences reflect both the sampled windows and the independently generated full-test corruption masks.

Across all 57 hard-protocol configurations, DiBiMa-Cardio-M achieved mean nMAE of 0.081 and mean PCC of 0.956, compared with 0.134 and 0.875 for DiBiMa-Cardio. It significantly improved nMAE, RMSE, SNR and PCC in 53 configurations, with a median nMAE ratio of 0.62. The four exceptions were ECG

denoising configurations, where nMAE increased by 5–14%. Under the standard protocol, DiBiMa-Cardio-M improved all seven reported metrics in all 57 configurations, with a median nMAE ratio of 0.67.

Downstream evaluations distinguish waveform fidelity from diagnostic information

Frozen representations were evaluated on PTB-XL diagnostic superclass classification (Fig. 6a,b). DiBiMa-Cardio-M achieved macro AUROC of 0.818 (95% CI 0.804–0.832), compared with 0.816 for DiBiMa-Cardio and 0.799–0.802 for randomly initialised counterparts. The difference between the pretrained models was 0.002 (95% CI −0.002 to 0.006), providing no evidence of improved clean-input classification after multimodal fine-tuning.

When classifiers trained on clean ECG were applied to gapped inputs, macro AUROC was 0.716 with DiBiMa-Cardio-M features versus 0.655 with DiBiMa-Cardio features, a difference of 0.061 (95% CI 0.049–0.072). Under hard noise, the values were 0.779 and 0.756. Restoring signals before feature extraction recovered much of the performance lost to corruption; for hard ECG imputation, AUROC reached 0.810 with DiBiMa-Cardio-M and 0.805 with DiBiMa-Cardio.

On MIMIC PERform AF, record-level cross-validation evaluated atrial-fibrillation classification across 35 recordings not used for backbone training (Fig. 6c). With the template prior zeroed, mean window accuracy for gapped ECG was 0.812 for DiBiMa-Cardio-M and 0.793 for DiBiMa-Cardio, increasing to 0.820 and 0.770 after model-based inpainting. For hard 2.0-s gaps without the prior, accuracy was 0.734 and 0.695. Classification from PPG-derived ECG yielded accuracies of 0.756 and 0.741, with no significant difference between models.

External translation fidelity depended strongly on direction (Fig. 6d). ECG-to-PPG translation improved with DiBiMa-Cardio-M: median per-record PCC increased from 0.599 to 0.722 in AF recordings and from 0.862 to 0.916 in non-AF recordings (paired Wilcoxon $p \leq 0.001$). PPG-to-ECG fidelity remained limited, with median PCC of approximately 0.21 in AF recordings and 0.28–0.31 in non-AF recordings, and no significant improvement between models ($p \geq 0.40$).

On the full cuff-less BP test set, PPG-to-ECG translation yielded median PCC of 0.696 for DiBiMa-Cardio-M versus 0.656 for DiBiMa-Cardio, with higher median PCC in every predicted diagnostic superclass (Fig. 6e). In the selected 209-window pool also evaluated with UniCardio-M, median PCC was 0.651, 0.646 and 0.468 for DiBiMa-Cardio-M, DiBiMa-Cardio and UniCardio-M, respectively ($p < 10^{-4}$ for DiBiMa-Cardio-M versus UniCardio-M).

Higher waveform correlation did not consistently preserve the diagnostic superclass assigned by the fixed ECG classifier. Predicted classes agreed between generated and recorded ECG in 46.9% of windows for DiBiMa-Cardio-M and 49.1% for DiBiMa-Cardio; myocardial infarction and hypertrophy classes were rarely preserved. These labels were model predictions rather than adjudicated diagnoses. Improved waveform similarity therefore did not establish preservation of diagnostic information.

Discussion

Cardiovascular waveform reconstruction involves a trade-off between recovering information from incomplete observations and the computational cost of producing an estimate. DiBiMa-Cardio shows that a single compact network, trained with diffusion-perturbed inputs, can cover denoising, imputation, forecasting, super-resolution and cross-modal translation across three signals in one deterministic forward pass. Its clearest advantage lies in accuracy per unit of computation: one network evaluation per window, rather than the 36 of the evaluated UniCardio-M configuration, gave lower mean nMAE across the 45 shared hard-protocol configurations at a fraction of the memory and processing time. The results also delimit where this formulation helps, and where reconstruction fidelity and preservation of diagnostic information diverge.

Diffusion perturbation during training and iterative sampling at inference affected reconstruction differently. The training comparison showed the clearest benefit in denoising, and direct inference achieved lower reconstruction error than the tested iterative sampler in most of the evaluated single-input configurations. Training on diffusion-perturbed inputs relates DiBiMa-Cardio to denoising autoencoders, which learn robust representations by reconstructing clean inputs from corrupted ones [24], and to generalised diffusion models that invert arbitrary degradations [25]. Unlike consistency models [18], it requires no dedicated one-step objective or teacher, and unlike iterative restoration schemes [13, 14, 26] it applies no refinement at inference. Within this model, the tested DDIM sampler did not improve nMAE, whereas diffusion training reduced error in most denoising configurations but increased it in most super-resolution configurations. The perception–distortion trade-off offers a consistent interpretation [27]. A deterministic estimator trained with a distortion loss approaches a central value of the conditional distribution of the target, which minimises distortion metrics such as nMAE but can attenuate fine detail that is uncertain given the input. This would favour one-pass reconstruction when the posterior is narrow, as in denoising and short-gap imputation, and disadvantage it when the posterior is broad, as in fourfold super-resolution and translation from a weakly informative source such as PPG to ECG. The same mechanism could contribute to the observation that closer waveform agreement did not consistently preserve predicted diagnostic classes. This account remains an interpretation, and distributional or perceptual metrics alongside nMAE would be needed to test it; the results do not show that iterative sampling is unnecessary for other diffusion formulations.

Relative to existing architectures, DiBiMa-Cardio combines four features that have not previously been brought together for cardiovascular signals (Supplementary Table 43). Pairwise translators address one source–target direction per model [9, 10, 28], general time-series imputers are not designed around pulsatile morphology [13, 14, 29], and the unified UniCardio retains iterative sampling [12]. DiBiMa-Cardio separates the training objective from inference; models time with bidirectional selective state-space scans [19, 20], which adapt to the input, unlike the time-invariant S4 layers of SSSD [14], and use context on both sides of a gap; covers super-resolution and forecasting with routing in one set of weights; and adds multimodal fusion through zero-initialised modules, so that DiBiMa-Cardio-M starts as

exactly the base model. Attention is retained in the encoders and for cross-modal fusion, so the network is not linear in sequence length overall.

Downstream analyses expose a boundary that waveform metrics alone do not reveal. Restoring corrupted ECG recovered much of the classification performance lost to missing data on PTB-XL [22], supporting restoration as a front-end that keeps automated analysis functional, including for representation models trained on clean recordings [30, 31]. In contrast, PPG-derived ECG showed weak correspondence on external AF recordings, and improved correlation did not consistently preserve predicted diagnostic classes, qualifying earlier demonstrations of cross-modal synthesis [9, 10]. A deterministic model returns a single plausible waveform even when the input is ambiguous. In arrhythmia this carries a specific risk: a gap imputed with regular beats could mask the irregularity that defines atrial fibrillation. This failure mode was not quantified here. Reconstructed segments should therefore be flagged as estimates, perturbation-based output bands should be validated for calibration before they are used as uncertainty, and reconstruction should be evaluated by the information it preserves for the intended task, not by waveform similarity alone.

The computational profile has practical consequences. With 5.36 million parameters, about 21 MB of single-precision weights and 92 MiB of peak GPU memory at batch size eight, DiBiMa-Cardio-M does not require data-centre hardware, and its amortised throughput implies that one workstation GPU could keep pace with many concurrent monitoring streams. Its strongest regime, single-input restoration, matches the wearable setting in which one optical sensor is available and most exposed to artefacts, and a model that accepts whichever signals are present suits emergency care and remote telemonitoring, where sensor availability changes [32]. These scenarios are motivated rather than demonstrated by the present measurements: all timing was obtained on one workstation GPU with fused Mamba-2 kernels, single-draw UniCardio-M was faster at batch size one, and processing an acquired window faster than its duration does not establish low-delay online operation.

Several limitations constrain these conclusions. Training and primary evaluation derive from one intensive-care source with window-level partitions, so patient-independent performance is not established; the benchmark subset contains 256 windows, and UniCardio-M was not evaluated on the full test set or retrained under the hard protocol. Per-window scaling precludes claims about BP in mmHg, the diffusion ablation was performed at smaller scale, and fusion stages lack matched training budgets. Patient-disjoint and multi-site evaluation, full-scale regression baselines, distributional metrics, quantified hallucination in arrhythmia and benchmarking on CPU and embedded hardware are the most direct next steps. Within these bounds, DiBiMa-Cardio indicates that conditional physiological reconstruction can be accurate, compact and inexpensive, provided that the information required by the downstream task is verified rather than assumed.

Methods

Data and preprocessing

All models were trained and evaluated on the Cuff-Less Blood Pressure Estimation dataset [33], which is derived from the MIMIC-II waveform database [34]. It contains 12,000 intensive-care recordings with synchronized fingertip PPG, invasive arterial BP and ECG sampled at 125 Hz. Preprocessing was applied per record. PPG and ECG were filtered with zero-phase fifth-order Butterworth filters (0.5-Hz high-pass and 50-Hz low-pass) and a 50-Hz notch ($Q = 30$); BP was left unfiltered. ECG polarity was rectified at record level with `neurokit2.ecg_invert` [35]. Records were then cut into non-overlapping 4-s windows (500 samples), giving 667,380 windows. BP was mapped to $(\mathrm{BP} - 100)/50$, and every channel was z-scored within each window. The model is therefore trained on waveform shape and relative amplitude, and absolute BP level is not modelled.

Windows were split at random (seed 42) into 627,380 training, 20,000 validation and 20,000 test windows. This is the split used by UniCardio [12]; the 256-window benchmark subset released with UniCardio matches 256 of our test windows exactly. The split is made at window level, so windows from the same record can fall in different partitions. The results therefore measure within-population generalization and are not subject-independent. Four test windows (one record) contained no finite ECG samples and were removed, leaving 19,996 test windows from 7,264 records (the "full test set").

Tasks and corruption protocols

Every setting is described by a target signal, a corruption and a set of auxiliary (extra) signals given in clean form. There are five task families. *Imputation* (I) recovers a contiguous missing segment, *forecasting* (F) predicts the final segment of the window, *denoising* (N) removes additive white Gaussian noise, and *super-resolution* (SR) recovers the 125-Hz signal from a 4× down-sampled copy. *Translation* (T) synthesizes a signal that was not recorded from one or two other signals. Each target $\in \{\mathrm{PPG}, \mathrm{ECG}, \mathrm{BP}\}$ is restored with no extra signal, with either of the other two signals, or with both, and there are six single-source and three dual-source translations. This gives 57 settings in total. Forty-five are supported by UniCardio-M, which has no SR route.

Two corruption protocols were used. The **hard protocol**, which is primary throughout the paper, uses a 2.0-s imputation gap centred on the stored gap position, a 2.0-s forecast horizon (the last 250 samples), and noise of standard deviation 0.71× the clean-window standard deviation (twice the training noise, about 3 dB SNR). The **standard benchmark protocol**, reported in the Supplementary Information, uses the gap and horizon masks stored with UniCardio's test subset (gap 1.20 s; horizon 1.00–1.33 s, median 1.17 s) and the training-level noise (0.355× the window standard deviation, about 9 dB SNR). The 2.0-s gap is clamped to the window, so in 43 of the 256 benchmark windows, where the stored gap lies near the end of the window, it coincides with the last 2.0 s and hard imputation receives the same input as hard forecasting. Corrupted inputs were generated once with a fixed seed and shared by all

models; for UniCardio-M the identical noise realization was rescaled to its own window normalization. Neither protocol matches every model's training distribution. DiBiMa-Cardio was trained with gaps of 0.67–1.33 s and horizons of 1.0–1.33 s, whereas the DiBiMa-Cardio-M fine-tuning (from the early-fusion stage on) used gaps of 1.0–2.5 s and horizons of 1.0–2.0 s. The 2.0-s gap and horizon therefore lie inside DiBiMa-Cardio-M's fine-tuning range but outside those of DiBiMa-Cardio and UniCardio-M, and the doubled noise lies outside the training range of every model. Stored masks exist only for the 256 benchmark windows. For the full test set, the standard corruption was therefore generated once with a fixed seed by the training data pipeline and shared by both models: one gap per window with a length in the training range (0.67–1.33 s), a horizon of 1.0–1.33 s and the same noise level. The hard protocol centres its 2.0-s gap on the same per-window gap position. We therefore report both protocols, together with continuous sweeps of noise scale ($\alpha \in [0, 3]$) and gap/horizon length (0.5–2.0 s).

DiBiMa-Cardio architecture

DiBiMa-Cardio (Fig. 2a) is an encoder–decoder network built from bidirectional Mamba-2 state-space blocks [19–21]. The input is a nine-channel tensor of length $L = 500$: three signal channels (PPG, ECG, BP; zero where unobserved), three binary observation masks, and three template-prior channels (described below).

Modality-specific encoders.

Each signal has its own condition encoder of 3 Bi-Mamba blocks with 4 layers each (model width 64, state size 16). A wired self-attention layer follows each block, and layer and batch normalization are interleaved. A bidirectional layer runs one Mamba-2 scan forward and one on the time-reversed sequence and sums the outputs. Each bidirectional state-space scan processes the sequence in linear time and provides context in both directions. The network also includes the attention operations described below.

Cross-modal attention.

The three latent sequences (3×64 channels) are concatenated along the time axis, and one multi-head attention block (8 heads; 16,768 parameters) attends over the joint $3L$ tokens. Absent modalities are masked as keys. This is the only place where DiBiMa-Cardio mixes modalities.

Decoders and routing.

Each target has its own Bi-Mamba decoder with a convolutional head. A route flag "source_target" picks which latent each decoder reads. Restoration uses self routes (PPG→PPG, ECG→ECG, BP→BP), and translation uses the source latent. Each decoder adds its raw input signal back to its latent (an internal residual). In total DiBiMa-Cardio has 3,870,899 parameters.

Template prior.

Following template-guided diffusion approaches for pulsatile-signal imputation [36], a beat template is estimated from the observed part of the corrupted target alone and

placed at the detected beat positions to fill its prior channel. It never uses the auxiliary signals. All other prior channels are zero. Training uses augmented templates, and evaluation a deterministic one ($K = 1$ template per window).

Diffusion training and one-step inference.

DiBiMa-Cardio is trained as a conditional diffusion model with $T = 20$ steps (scaled-linear $\beta \in [10^{-5}, 2 \times 10^{-2}]$). It predicts the clean signal (x_0 parameterization) from inputs perturbed by the forward process $q(x_t \mid x_0)$ at a random step t, with a sinusoidal step embedding (`time_mlp`). In contrast to latent or target-slot diffusion, noise is applied to the observed input channels, so the model learns to restore from a perturbed view of what it can see. At inference we run a single forward pass at $t = T - 1$ with unperturbed inputs, which gives a deterministic estimate at the cost of one network evaluation. Iterative DDIM sampling [17] with DiBiMa-Cardio did not improve accuracy (Supplementary Information), so all results are one-step. BiMa denotes the same network trained without the diffusion perturbation and step embedding.

Training.

DiBiMa-Cardio was trained on the full training split for 5 epochs (batch size 32, AdamW [37] with weight decay 10^{-6}, learning rate decayed linearly from 10^{-3} to 10^{-4} per step). It used modality sampling, in which each step draws the corrupted modalities and task, and a multi-task loss (MSE over the scored region, with task-specific weights) over imputation, forecasting, denoising, super-resolution and translation. Training corruptions were gaps $\sim \mathcal{U}(0.67, 1.33)$ s, horizons $\sim \mathcal{U}(1.0, 1.33)$ s, white Gaussian noise ($0.355\times$ the window standard deviation) and $4\times$ down-sampling. In this pre-training stage all three modalities are corrupted jointly in restoration tasks, so DiBiMa-Cardio never saw a clean auxiliary signal next to a corrupted target.

DiBiMa-Cardio-M: multimodal fine-tuning

DiBiMa-Cardio-M (Fig. 2b) starts from DiBiMa-Cardio and is fine-tuned on all 57 evaluation settings plus four joint-replay tasks, so that restoration uses clean auxiliary signals when they are present. Four modules were added, each zero-initialized so that the network is exactly DiBiMa-Cardio at step 0. (i) A *gated Bi-Mamba bottleneck* (1 block $\times$ 1 layer over the 192-channel concatenated latent, with a per-channel output gate; 1.02 M parameters) adds a time-aligned fusion path between encoders and decoders. (ii) *Early-fusion* convolutions (2,816 parameters) pass the auxiliary signals into each encoder's input. (iii) *Decoder-fusion adapters* (kernel 9, 64 hidden channels; 467,712 parameters) inject the availability-gated features of the other two modalities at the decoder input and after the decoder Bi-Mamba. (iv) *Modality- and state-aware cross-attention* (392 parameters) adds modality and observed/missing/absent-state embeddings to queries and keys, together with a gated rotary time encoding [38]. DiBiMa-Cardio-M has 5,360,351 parameters.

The stages were trained in sequence. Each stage starts from the best checkpoint of the previous one, selected every 1,000 steps on 1,024 validation windows (multimodal tasks only for the decoder-fusion and aware-attention stages): gated bottleneck (5,000 steps selected); + early fusion, with gaps 1.0–2.5 s and horizons 1.0–2.0 s (4,000); +

decoder fusion, with batch-norm statistics frozen and translation tasks up-weighted $\times 2$ (20,000); + aware cross-attention (14,000); and a final stage of 10,000 steps with *template-prior dropout*, in which the prior is zeroed with probability 0.5 in imputation and forecasting windows that have clean auxiliary signals. This forces the model to read the gap from the auxiliary signals rather than from the target's own template; evaluation keeps the prior. Every stage used 8 tasks per step with one backward pass per task, 500 warm-up steps followed by cosine decay, weight decay 10^{-6}, and a mixed input in which half of each batch keeps clean inputs and half receives q-sampled inputs. The fusion stages used batch size 32 and learning rates of 3×10^{-5} for pretrained weights and 3×10^{-4} for new modules; multimodal tasks were up-weighted $\times 2$ in task sampling in the bottleneck stage and $\times 3$ from early fusion onward. A parallel fine-tune of DiBiMa-Cardio with the same tasks but no new module (the "no bottleneck" control in Fig. 3) gauges the effect of fine-tuning alone. It used batch size 64, a single learning rate of 10^{-4} and $\times 2$ multimodal weighting (best checkpoint at 1,000 steps), so it is not hyperparameter-matched to the bottleneck stage. The stages were not given equal-step controls, so the trajectory in Fig. 3 is a cumulative ablation and not an attribution of the gain to each module.

UniCardio-M baseline

UniCardio [12] is a multimodal diffusion transformer (8.12 M parameters) that generates a target signal from any combination of PPG, ECG and BP conditions. We used the released multimodal checkpoint with 6-step DDIM sampling and refer to it as *UniCardio-M* throughout. The released code provides DDIM only for single-condition routes, so we implemented the same sampler for multi-condition routes. On the eight single-input routes it supports, our implementation is bit-identical to the released one. UniCardio-M received its own preprocessing (per-window min–max scaling to $[-1, 1]$ for PPG and ECG, and $(\mathrm{BP} - 100)/50$ for BP) together with the masks and noise realizations described above. UniCardio-M has no super-resolution route and was not trained for forecasting; its forecasting outputs use the imputation route with a terminal mask. UniCardio's original DDPM-50 sampler is 8–9 times slower than DDIM-6 (Supplementary Information).

UniCardio original training protocol.

UniCardio [12] is trained on one fixed corruption per task. In the released training script, the denoising inputs are made by adding white Gaussian noise at a fixed signal-to-noise ratio of 15 dB. The imputation inputs are made by masking one contiguous block of 30% of the 4-s window (1.2 s at 125 Hz), at a random position. Both corrupted copies are generated once, before training, and reused at every epoch. The released test set follows the same recipe: a 1.2-s gap and a forecast horizon of 1.00–1.33 s. The model therefore never sees other noise levels, gap lengths or horizons, and its error grows when the test corruption moves slightly away from this point in either direction. In the single-input gap sweep (Supplementary Fig. 6), its PPG imputation error is lowest at 1.0 s (min–max MAE 0.082). It rises for shorter gaps (0.097 at 0.5 s) as well as for longer ones (0.304 at 2.0 s). On noise-free input, equivalent to a simple reconstruction task, its denoising route degrades the signal, a case it was not trained for.

We nevertheless evaluate the released checkpoint unchanged, for three reasons. First, the corruption distribution is part of the method. Retraining UniCardio with randomized noise levels and gap lengths would give a different model, built on choices (ranges, sampling, schedule) that the original authors did not design or validate. Any shortfall could then be blamed on our reimplementation rather than on the architecture, whereas the authors' weights keep the baseline in its published state. Second, retraining from scratch is costly: the released configuration runs 800 epochs at batch size 256 and is set up for seven GPUs. A fair retraining would also need its own hyperparameter search. Third, our comparison does not rest on a single operating point. We report UniCardio's native benchmark corruption alongside the hard protocol, and we sweep the noise level and the gap/horizon length continuously. The reader can therefore see where UniCardio's advantage holds (for example, multimodal imputation with clean auxiliary signals at gaps of 1–2 s) and where it does not. For transparency, the hard protocol is not equally out of distribution for all models. Gaps and horizons of 2 s fall inside the DiBiMa-M fine-tuning ranges (gaps 1.0–2.5 s, horizons 1.0–2.0 s). They fall outside the training ranges of UniCardio and of base DiBiMa (gaps 0.67–1.33 s, horizons 1.0–1.33 s). Twice the standard test noise is outside the training distribution of every model. Robustness to an unknown corruption level is thus part of what we compare: DiBiMa-M was trained for it, UniCardio was not.

Metrics and statistics

Following UniCardio, the primary error is the *normalized MAE* (nMAE). Prediction and reference are each min–max scaled to $[-1, 1]$ within the window, and the MAE is computed over the scored region: the gap for imputation, the horizon for forecasting, and the whole window otherwise. We also report the Pearson correlation coefficient (PCC), MSE, RMSE, SNR ($10 \log_{10} \|y\|^2 / \|y - \hat{y}\|^2$), PSNR and the two-sample Kolmogorov–Smirnov statistic between predicted and reference amplitude distributions. Each metric is computed both on the min–max scale and on a reference-z scale (both signals standardized with the reference window's mean and standard deviation). Each model is scored against its own preprocessed reference, so the min–max nMAE is the only error directly comparable across models.

Per-setting means carry 95% confidence intervals from a record-cluster bootstrap (2,000 resamples, seed 417). Paired model comparisons on the same windows use a record-clustered sign-flip permutation test on per-window differences (10,000 resamples, seed 418), with Benjamini–Hochberg correction [39] over all settings in a comparison. A setting counts as a win or a loss only when $q < 0.05$; otherwise it is a tie. Results on the 256-window subset (where UniCardio-M was run) and on the full 19,996-window test set (DiBiMa-Cardio and DiBiMa-Cardio-M only) are reported separately.

Efficiency

Latency and memory were measured on one NVIDIA RTX 4000 Ada GPU (20 GB; PyTorch 2.9, CUDA 12.8, fp32 with TF32 matrix multiplication, fused Mamba-2

kernels) with one other process on the GPU (about 0.2 GB and 17–23% utilization during the run). Four settings were timed (ECG-I, ECG-I+PPG+BP, PPG→ECG and PPG+BP→ECG) with hard-protocol inputs, at batch sizes of 1, 8, 32 and 64. For each configuration, three warm-up batches were discarded and the CUDA-synchronized wall time per batch was recorded over 30 repeats (15 at batch $\geq$ 32; 10 and 5 for UniCardio-M; 3 for DDPM-50). We report the median time per window and the peak allocated GPU memory, including weights. UniCardio-M was timed as evaluated (6-step DDIM, mean of 6 draws, 36 network evaluations per window). It was also timed with a single draw (6 evaluations) as a speed reference whose accuracy we did not measure, and with its original DDPM-50 sampler (6 draws, 300 evaluations). DiBiMa-Cardio and DiBiMa-Cardio-M run one evaluation per window. The Mamba-2 kernels require a GPU, so no CPU timing is reported.

Downstream tasks

Downstream evaluations used frozen DiBiMa-Cardio and DiBiMa-Cardio-M under the same inference contract as above. Each experiment corrupted the inputs with a fixed random generator shared by both models, under a standard condition (noise 0.355× the window standard deviation; one gap of 0.67–1.33 s, or the 1.2-s gap for MIMIC) and a hard condition (the same noise realization doubled; a 2.0-s gap centred on the standard gap).

PTB-XL ECG classification.

PTB-XL [22] lead II was used with the official record split (17,084 training, 2,146 validation and 2,158 test records). Each record was resampled to 125 Hz, filtered as in training, split into two 5-s halves and z-scored. For each record we pooled the ECG block of the model latent (64 channels, after the encoders and, for DiBiMa-Cardio-M, the bottleneck) by mean, standard deviation and maximum over time, giving a 192-dimensional feature. A two-layer MLP head (256–128 units, GELU, dropout 0.3, AdamW, early stopping on validation macro AUROC) was trained for the five diagnostic superclasses (NORM, MI, STTC, CD, HYP) with five seeds, and predictions were averaged over the two halves and over the seeds. Two head-training regimes were used. *Matched* heads were trained and tested on the same input condition (clean, noisy or gapped ECG). *Clean-trained* heads were trained on clean ECG and tested on corrupted or model-restored ECG, which measures representation shift under corruption. Randomly initialized networks of each architecture served as controls. Macro AUROC carries 95% confidence intervals and paired contrasts from 2,000 test-record bootstrap resamples.

Atrial-fibrillation detection on MIMIC PERform.

The MIMIC PERform AF dataset [23] contains 35 records (19 AF, 16 non-AF) of PPG and ECG at 125 Hz, which we cut into 10,355 4-s windows. This dataset was not used for training. The classifier inputs were frozen-latent features of the PPG, of the ECG generated from PPG, of the real ECG, of the gapped ECG and of the model-inpainted ECG, with the same head as above. The primary protocol was stratified five-fold cross-validation at record level, repeated five times, with 2,000 record-bootstrap resamples.

UniCardio's window-level ten-fold protocol, which places windows of one record in both training and test folds, is reported as secondary. For gapped ECG the template prior encodes beat positions detected in the observed ECG and can leak rhythm information to the classifier, so the leak-free comparison zeroes the prior channels. Translation fidelity (PCC and nMAE per record, PPG→ECG and ECG→PPG) was compared with paired Wilcoxon tests.

PPG-to-ECG fidelity by diagnostic class.

On the 19,996 full-test windows, ECG translated from PPG was compared with the recorded ECG. Each window was assigned the superclass predicted by a fixed labeller (the DiBiMa-Cardio clean PTB-XL head applied to the recorded ECG). These labels are model predictions, not clinical diagnoses. We report the median PCC per class with record-cluster bootstrap intervals, and class agreement, defined as the fraction of windows whose generated ECG received the same predicted class as the recorded ECG. UniCardio-M was compared on a pool of 209 windows, selected per class as the top 20% by labeller confidence, with one window per record and up to 50 per class, restricted to windows whose raw PPG could be located in UniCardio's data ($r > 0.9$).

Interactive demonstration

A public web application (https://huggingface.co/spaces/Ugo96/dibima-cardio) runs DiBiMa-Cardio, DiBiMa-Cardio-M and UniCardio-M live on the 256 benchmark windows. The user selects the input signals, the task, the corruption (hard, standard or custom gap, horizon and noise scale) and the windows, and the app reports waveforms, metrics and measured inference times.

Data availability. All datasets are public. The Cuff-Less Blood Pressure Estimation dataset [33] is available from the UCI Machine Learning Repository, PTB-XL [22] from PhysioNet, and the MIMIC PERform AF dataset [23] from its public repository. The 256-window benchmark subset is the one released with UniCardio [12]. Source data for every figure are provided with this paper.

Code availability. Code for training, fine-tuning, evaluation and figure generation is available at https://github.com/UgoLomoio/DiBiMa-Cardio. An interactive demonstration running DiBiMa-Cardio, DiBiMa-Cardio-M and UniCardio-M live is available at https://huggingface.co/spaces/Ugo96/dibima-cardio.

Supplementary information. The Supplementary Information contains Supplementary Notes 1–3 (corruption protocols, one-step sampling, weight-space programming experiments), Supplementary Figs. 1–16 and Supplementary Tables 1–43, covering every task and input combination under both corruption protocols, the full-test-set metrics and the downstream details.

Author contributions. U.L. wrote the first draft, developed the code, performed the experiments, and contributed to the methodology, formal analysis, and investigation. P.H.G., P.V. and P.L. contributed to conceptualization and supervision. P.H.G. contributed to the methodology and formal analysis. All authors contributed to the investigation, revised the manuscript, and approved the final version.

Competing interests. The authors declare no competing interests.

References

[1] Mukkamala, R. *et al.* Toward ubiquitous blood pressure monitoring via pulse transit time: Theory and practice. *IEEE Transactions on Biomedical Engineering* **62**, 1879–1901 (2015).

[2] Finnegan, E. *et al.* Features from the photoplethysmogram and the electrocardiogram for estimating changes in blood pressure. *Scientific Reports* **13**, 986 (2023).

[3] Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C. & Nazeran, H. A review on wearable photoplethysmography sensors and their potential future applications in health care. *International Journal of Biosensors & Bioelectronics* **4**, 195–202 (2018).

[4] Lee, J., Kim, M., Park, H.-K. & Kim, I. Y. Motion artifact reduction in wearable photoplethysmography based on multi-channel sensors with multiple wavelengths. *Sensors* **20**, 1493 (2020).

[5] Nardelli, M. & Bailón, R. Advances in wearable photoplethysmography applications in health monitoring. *Sensors* **23**, 7064 (2023).

[6] Chen, G., Zou, L. & Ji, Z. A review: Blood pressure monitoring based on ppg and circadian rhythm. *APL Bioengineering* **8**, 031501 (2024).

[7] Patel, S., Park, H., Bonato, P., Chan, L. & Rodgers, M. A review of wearable sensors and systems with application in rehabilitation. *Journal of NeuroEngineering and Rehabilitation* **9**, 21 (2012).

[8] Olmedo-Aguirre, J. O. *et al.* Remote healthcare for elderly people using wearables: A review. *Biosensors* **12**, 73 (2022).

[9] Sarkar, P. & Etemad, A. Cardiogan: Attentive generative adversarial network with dual discriminators for synthesis of ecg from ppg. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**, 488–496 (2021).

[10] Shome, D., Sarkar, P. & Etemad, A. Region-disentangled diffusion model for high-fidelity ppg-to-ecg translation. *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence* (2024). URL https://doi.org/10.1609/aaai.v38i13.29422.

[11] Pan, Y. *et al.* Hipno: Symmetry-aware physics-informed neural operators for noninvasive hemodynamic inference (2026). URL https://arxiv.org/abs/2608.10011. arXiv:2608.10011.

[12] Chen, Z. *et al.* Versatile cardiovascular signal generation with a unified diffusion transformer. *Nature Machine Intelligence* **8**, 6–19 (2026).

[13] Tashiro, Y., Song, J., Song, Y. & Ermon, S. CSDI: conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems* **34**, 24804–24816 (2021).

[14] Lopez Alcaraz, J. M. & Strodthoff, N. Diffusion-based time series imputation and forecasting with structured state space models. *Transactions on Machine Learning Research* (2023). URL https://arxiv.org/abs/2208.09399.

[15] Lopez Alcaraz, J. M. & Strodthoff, N. Diffusion-based conditional ecg generation with structured state space models. *Computers in Biology and Medicine* **163**, 107115 (2023).

[16] Li, X., Sakevych, M., Atkinson, G. & Metsis, V. Biodiffusion: A versatile diffusion model for biomedical signal synthesis. *Bioengineering* **11**, 299 (2024).

[17] Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. *International Conference on Learning Representations* (2021).

[18] Song, Y., Dhariwal, P., Chen, M. & Sutskever, I. Krause, A. *et al.* (eds) *Consistency models.* (eds Krause, A. *et al.*) *Proceedings of the 40th International Conference on Machine Learning*, Vol. 202 of *Proceedings of Machine Learning Research*, 32211–32252 (PMLR, 2023). URL https://proceedings.mlr.press/v202/song23a.html.

[19] Gu, A. & Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* (2023).

[20] Dao, T. & Gu, A. Transformers are SSMs: generalized models and efficient algorithms through structured state space duality. *Proceedings of the 41st International Conference on Machine Learning* **235**, 10041–10071 (2024).

[21] Lomoio, U., Lió, P., Guzzi, P. H. & Veltri, P. Bidirectional mamba-2 boosts eeg super-resolution via regression and diffusion. *Bioinformatics* btag169 (2026).

[22] Wagner, P. *et al.* PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data* **7**, 154 (2020).

[23] Charlton, P. H. *et al.* Detecting beats in the photoplethysmogram: benchmarking open-source algorithms. *Physiological Measurement* **43**, 085007 (2022).

[24] Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders 1096–1103 (2008).

[25] Bansal, A. *et al.* Cold diffusion: Inverting arbitrary image transforms without noise **36** (2023).

[26] Delbracio, M. & Milanfar, P. Inversion by direct iteration: An alternative to denoising diffusion for image restoration. *Transactions on Machine Learning Research* (2023). URL https://arxiv.org/abs/2303.11435.

[27] Blau, Y. & Michaeli, T. The perception-distortion tradeoff 6228–6237 (2018).

[28] Ibtehaz, N. *et al.* Ppg2abp: Translating photoplethysmogram (ppg) signals to arterial blood pressure (abp) waveforms. *Bioengineering* **9** (2022). URL https://www.mdpi.com/2306-5354/9/11/692.

[29] Xu, M. *et al.* Pulseimpute: A novel benchmark task for pulsative physiological signal imputation. *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (2022). URL https://openreview.net/forum?id=x_kBZYiUrxR.

[30] Moody, J. B. *et al.* A foundation transformer model with self-supervised learning for ecg-based assessment of cardiac and coronary function. *NEJM AI* **2** (2025).

[31] Jiang, H. *et al.* Ecg-mamba: Cardiac abnormality classification with non-uniform state space models. *arXiv preprint arXiv:2510.13046* (2025).

[32] Koehler, F., Koehler, K., Deckwart, O. *et al.* Efficacy of telemedical interventional management in patients with heart failure (TIM-HF2): a randomised, controlled, parallel-group, unmasked trial. *The Lancet* **392**, 1047–1057 (2018).

[33] Kachuee, M., Kiani, M. M., Mohammadzade, H. & Shabany, M. Cuff-less high-accuracy calibration-free blood pressure estimation using pulse transit time. *2015 IEEE International Symposium on Circuits and Systems (ISCAS)* 1006–1009 (2015).

[34] Saeed, M. *et al.* Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. *Critical Care Medicine* **39**, 952–960 (2011).

[35] Makowski, D. *et al.* NeuroKit2: a Python toolbox for neurophysiological signal processing. *Behavior Research Methods* **53**, 1689–1696 (2021).

[36] Jenkins, A. & Chen, Y. Improving diffusion models for ecg imputation with an augmented template prior. *arXiv preprint arXiv:2310.15742* (2023). URL https://arxiv.org/abs/2310.15742.

[37] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. *International Conference on Learning Representations* (2019).

[38] Su, J. *et al.* RoFormer: enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024).

[39] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

Data availability

All datasets are public. The Cuff-Less Blood Pressure Estimation dataset [33] is available from the UCI Machine Learning Repository, PTB-XL [22] from PhysioNet, and the MIMIC PERform AF dataset [23] from its public repository. The 256-window benchmark subset is the one released with UniCardio [12]. Source data for every figure are provided with this paper.

Code availability

Code for training, fine-tuning, evaluation and figure generation is available at https://github.com/UgoLomoio/DiBiMa-Cardio. An interactive demonstration running DiBiMa-Cardio, DiBiMa-Cardio-M and UniCardio-M live is available at https://huggingface.co/spaces/Ugo96/dibima-cardio.

Author contributions

P.H.G., P.L. and U.L. wrote the first draft of the paper. All authors commented on previous versions of the paper, and all authors read and approved the final paper. P.H.G. and P.L. worked on conceptualization and supervision. P.H.G. and U.L. worked on the methodology and formal analysis. All authors contributed to the investigation.

Competing interests

The authors declare no competing interests.

Figure legends

Figure 1. Cardiovascular signal reconstruction under real-world monitoring conditions. a, ECG, PPG and arterial BP give complementary views of cardiovascular function: electrical activation, peripheral pulse dynamics and haemodynamic load [1, 2]. Their joint analysis links timing, morphology and pressure, for example through the pulse transit time (PTT) between the ECG R peak and the PPG pulse foot. **b**, In practice, motion artefacts, poor sensor contact and low perfusion distort recordings, interruptions leave missing segments, and invasive BP is unavailable in many monitoring settings [3–6]. **c**, Each problem defines a reconstruction task: denoising (N), imputation (I), forecasting (F), fourfold super-resolution (SR) and cross-modal translation. Any of the three signals can be the target, with zero, one or two clean auxiliary signals available. Black, observed input; green, reconstructed target; orange shading, corrupted or unobserved region. **d**, The model uses the available waveforms and observation masks to estimate the target waveform. For example, DiBiMa-Cardio-M can estimate a missing ECG segment from the surrounding ECG samples and a simultaneously recorded clean PPG waveform. A template prior derived from the observed target provides additional conditioning for imputation and forecasting. Training perturbs observed inputs with diffusion noise, and inference reconstructs the target in one deterministic pass. Figure 2 details the architecture and the multimodal fusion modules. Waveforms are schematic illustrations, not model outputs.

Figure 2. Unified one-pass cardiovascular reconstruction. a, DiBiMa-Cardio combines modality-specific bidirectional Mamba-2 encoders, cross-modal attention and routed decoders. Inputs comprise three signals, observation masks and template priors (500 samples; 4 s at 125 Hz). Diffusion training perturbs observed inputs; inference uses one deterministic pass. **b**, DiBiMa-Cardio-M adds multimodal fusion modules and template-prior dropout, increasing the parameter count from 3.87 to 5.36 million. **c**, The 57 configurations cover imputation (I), denoising (N), forecasting (F), fourfold super-resolution (SR) and translation with different input combinations. Hatched cells indicate 12 SR configurations unsupported by UniCardio-M. **d**, Standard and hard corruption protocols: gaps of 1.20 and 2.0 s, forecast horizons of 1.00–1.33 and 2.0 s, and noise scales $\alpha = 1$ and 2, respectively; $\alpha = 1$ corresponds to 0.355 times the signal standard deviation. **e**, Training and evaluation workflow. All three models are compared on 256 windows from 247 records; full-test evaluation includes 19,996 windows from 7,264 records for the DiBiMa models.

Figure 3. Effects of diffusion training, inference and multimodal fine-tuning. a, Matched smaller-scale DiBiMa-Cardio and non-diffusion BiMa models, evaluated on 226 windows from 217 records outside training and validation under the hard protocol. Bars show significant improvements, nonsignificant differences and deteriorations; ratio summaries show medians and ranges. Diffusion training improves denoising but impairs SR. **b**, One-pass inference versus six-step DDIM averaged over six draws across 18 single-input configurations under the standard protocol. Bars show task means, dots individual configurations and numbers significant one-pass improvements. Subset-run median processing times are 16.8 versus 534 ms per window at batch size eight; whiskers indicate interquartile ranges. **c**, Median nMAE ratios relative to UniCardio-M along the fusion-development trajectory (45 configurations; hard protocol). Stages have unequal training budgets and do not isolate module contributions. **d**, DiBiMa-Cardio-M versus DiBiMa-Cardio across 57 hard-protocol configurations: nMAE percentage changes and PCC differences, with paired record-bootstrap 95% confidence intervals where shown. Colours indicate record-clustered sign-flip tests with Benjamini–Hochberg correction ($q < 0.05$); nonsignificant differences do not establish equivalence. nMAE is MAE after independent per-window min–max scaling to $[-1, 1]$ over the scored region.

Figure 4. Reconstruction accuracy and robustness relative to UniCardio-M. Comparisons use 256 benchmark windows from 247 records under the hard protocol (2.0-s gaps and horizons; doubled noise). UniCardio-M uses six-step DDIM averaged over six draws. **a,b**, Mean nMAE (lower is better) and PCC (higher is better), with record-cluster bootstrap 95% confidence intervals (2,000 resamples). Rows indicate target modalities; columns indicate tasks and auxiliary signals (+E, ECG; +P, PPG; +B, BP) or translation sources. Green and grey asterisks favour DiBiMa-Cardio-M and UniCardio-M, respectively (paired record-clustered sign-flip tests; Benjamini–Hochberg $q < 0.05$ across 45 shared configurations). ns denotes no significant difference. Crosses indicate unsupported SR routes; daggers indicate significant DiBiMa-Cardio-M improvements over DiBiMa-Cardio across the 12 SR configurations. **c**, Significant improvements/nonsignificant differences/deteriorations: 22/14/9 for nMAE and 29/10/6 for PCC. **d**, Mean nMAE across corruption sweeps, without auxiliary signals (top; three configurations) or with them (bottom; nine). Grey bands or dotted lines mark UniCardio-M benchmark corruption levels; orange dashed lines mark the hard protocol. Standard-protocol and per-configuration results are provided in the Supplementary Information.

Figure 5. Representative reconstructions and computational efficiency. a, Hard-protocol examples selected to approximate median error ranks across models. Traces are independently min–max scaled to $[-1, 1]$; shading identifies the scored gap or horizon, and coloured numbers give per-window nMAE. Black, recorded signal; green, DiBiMa-Cardio-M; grey, UniCardio-M; dashed blue, DiBiMa-Cardio where shown. UniCardio-M was rerun on selected windows, so its stochastic outputs differ from the full benchmark run. **b,c**, Peak allocated GPU memory, including weights, and amortised processing time per 4-s window on an NVIDIA RTX 4000 Ada. Points summarise medians over four settings; latency whiskers span their minimum and maximum. The evaluated UniCardio-M configuration uses six-step DDIM and six draws (36 network evaluations); single-draw DDIM and DDPM-50 curves provide timing references only. Single-draw accuracy was not evaluated. The horizontal dashed line in **c** marks the 4-s window duration. **d**, Mean nMAE across 45 shared hard-protocol configurations versus processing time at batch sizes one (open markers) and eight (filled). Detailed timing procedures and measurements are given in Methods and Supplementary Table 31.

Figure 6. Downstream classification and cross-modal fidelity. a, PTB-XL lead-II classification across five diagnostic superclasses (2,158 test records). Bars show macro AUROC for five-seed classifier ensembles trained on frozen features; hatching indicates randomly initialised backbones. Whiskers show record-bootstrap 95% confidence intervals (2,000 resamples). **b**, Clean-trained classifiers evaluated on corrupted inputs (light bars) or model-restored inputs (dark bars); dashed lines indicate clean-input AUROC. Asterisks compare models on corrupted inputs using paired record bootstraps ($p < 0.05$). **c**, AF detection on 35 MIMIC PERform records (10,355 windows), using record-level fivefold cross-validation repeated five times. Bars show mean window accuracy; whiskers show standard deviations across repeats. Gapped-input comparisons zero the template prior; I denotes imputation and hard denotes 2.0-s gaps. Asterisks indicate uncorrected paired record-bootstrap tests on repeat-averaged out-of-fold probabilities. **d**, External translation fidelity: dots show per-record mean PCC, paired lines connect models and black bars show medians (19 AF; 16 non-AF records). Tests are paired Wilcoxon signed-rank comparisons. **e**, PPG-to-ECG median PCC by classifier-predicted superclass, with record-cluster bootstrap 95% confidence intervals: all 19,996 test windows (left) and the selected 209-window UniCardio-M comparison pool (right). Asterisks indicate paired Wilcoxon tests on record means (left) or selected windows (right). Labels are model predictions, not clinical diagnoses; ns indicates no significant difference.

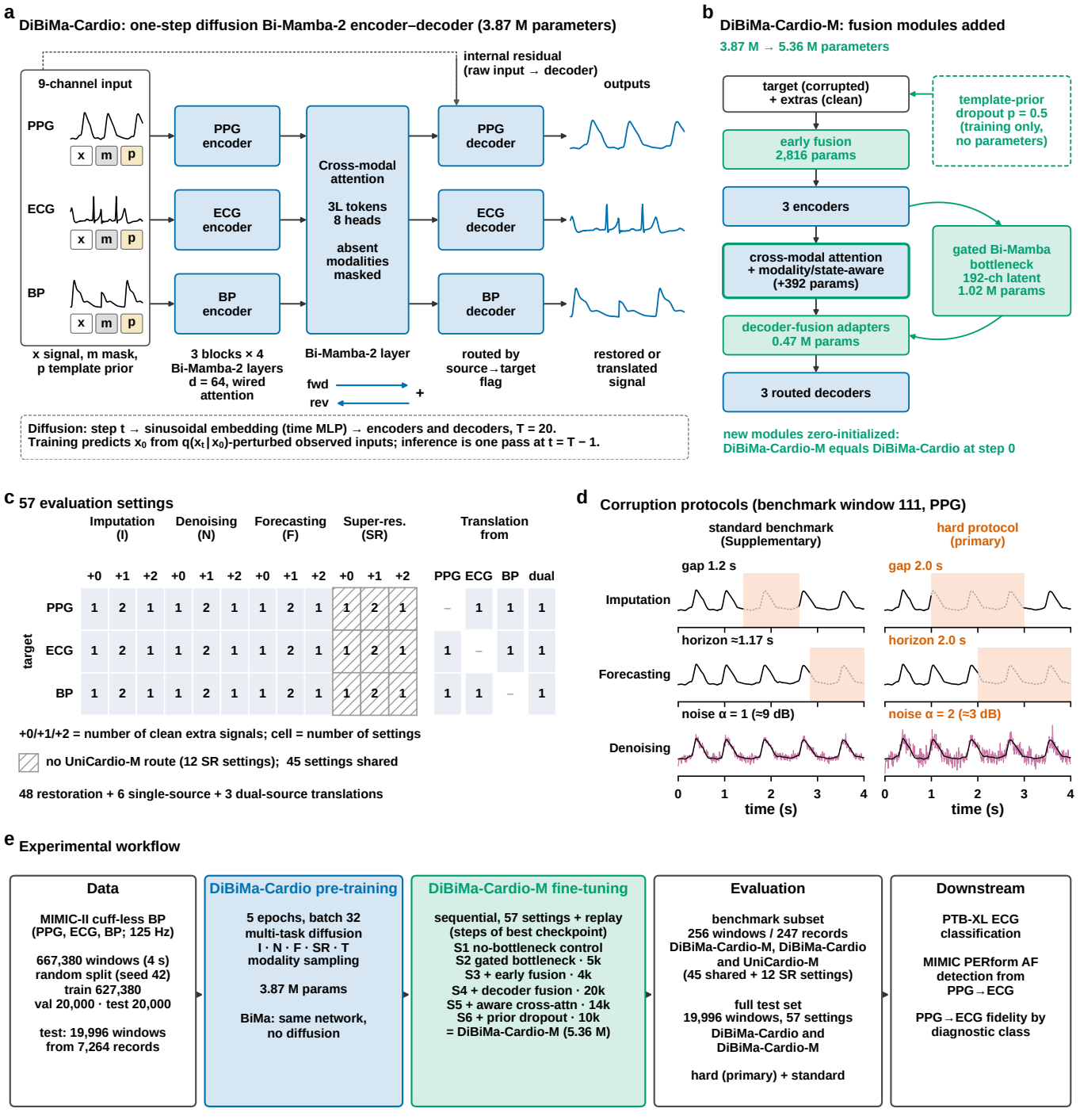


Fig. 2 Unified one-pass cardiovascular reconstruction. Detailed panel descriptions and statistical definitions are provided in Figure legends.

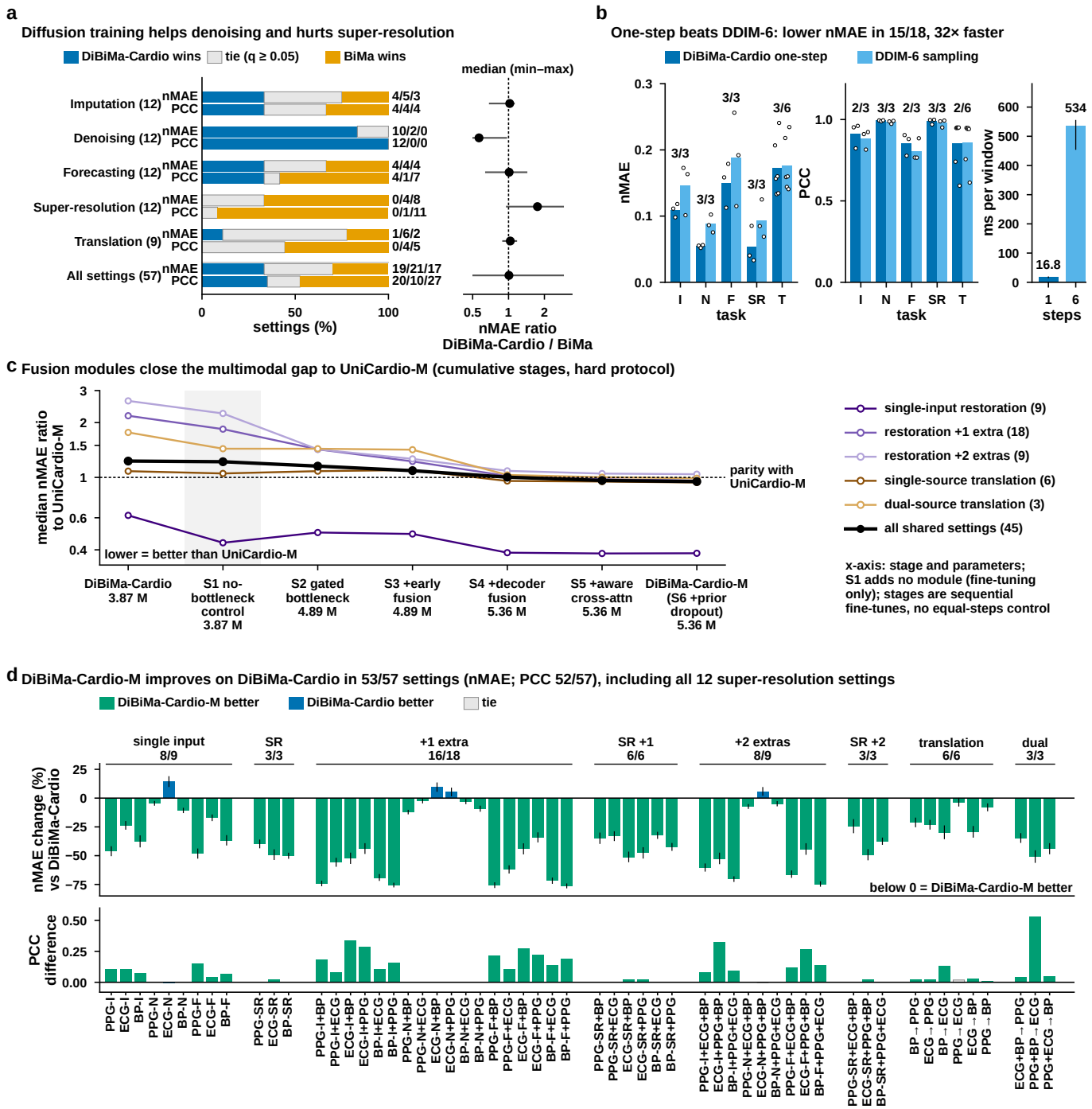


Fig. 3 Effects of diffusion training, inference and multimodal fine-tuning. Detailed panel descriptions and statistical definitions are provided in Figure legends.

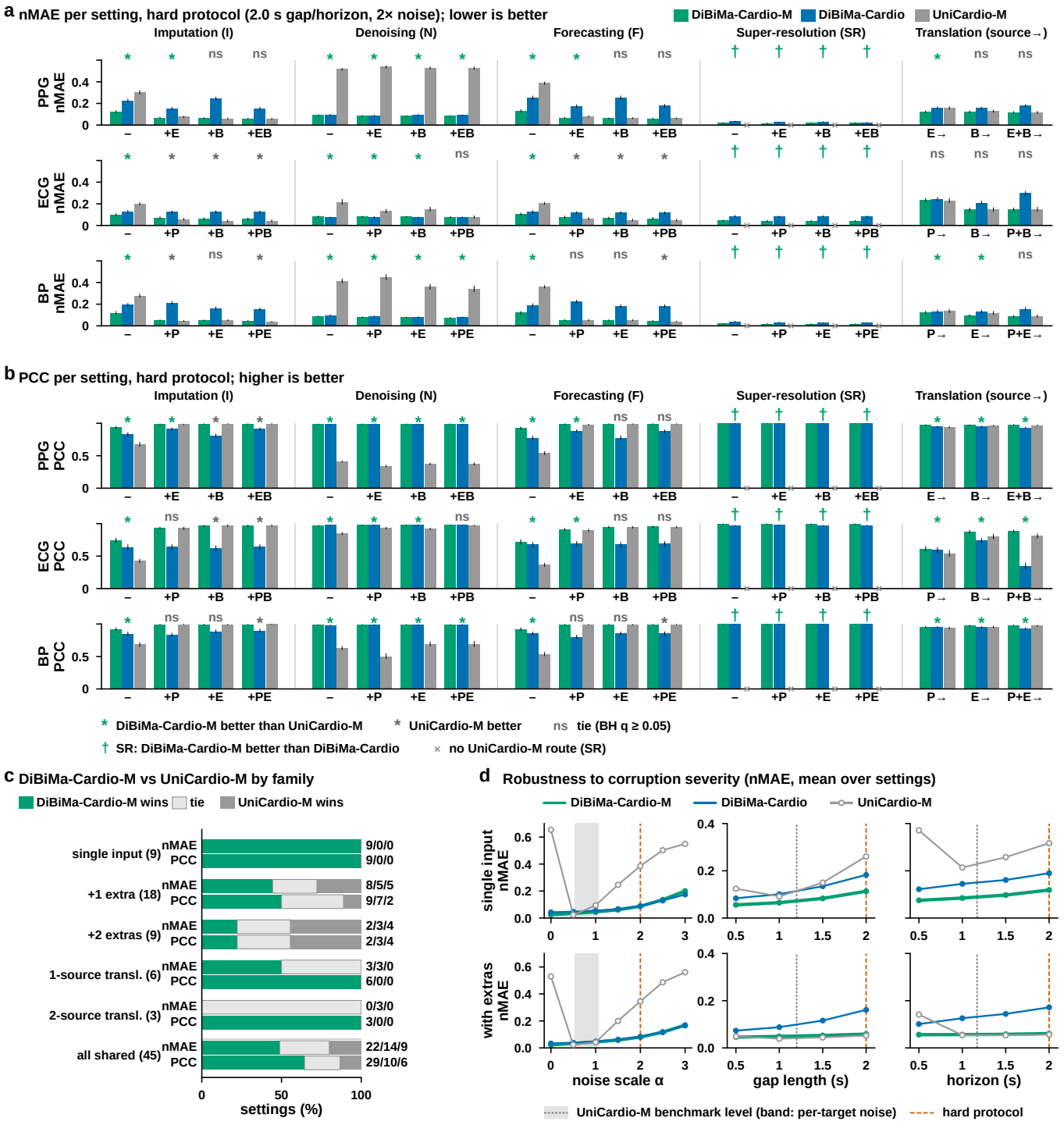


Fig. 4 Reconstruction accuracy and robustness relative to UniCardio-M. Detailed panel descriptions and statistical definitions are provided in Figure legends.

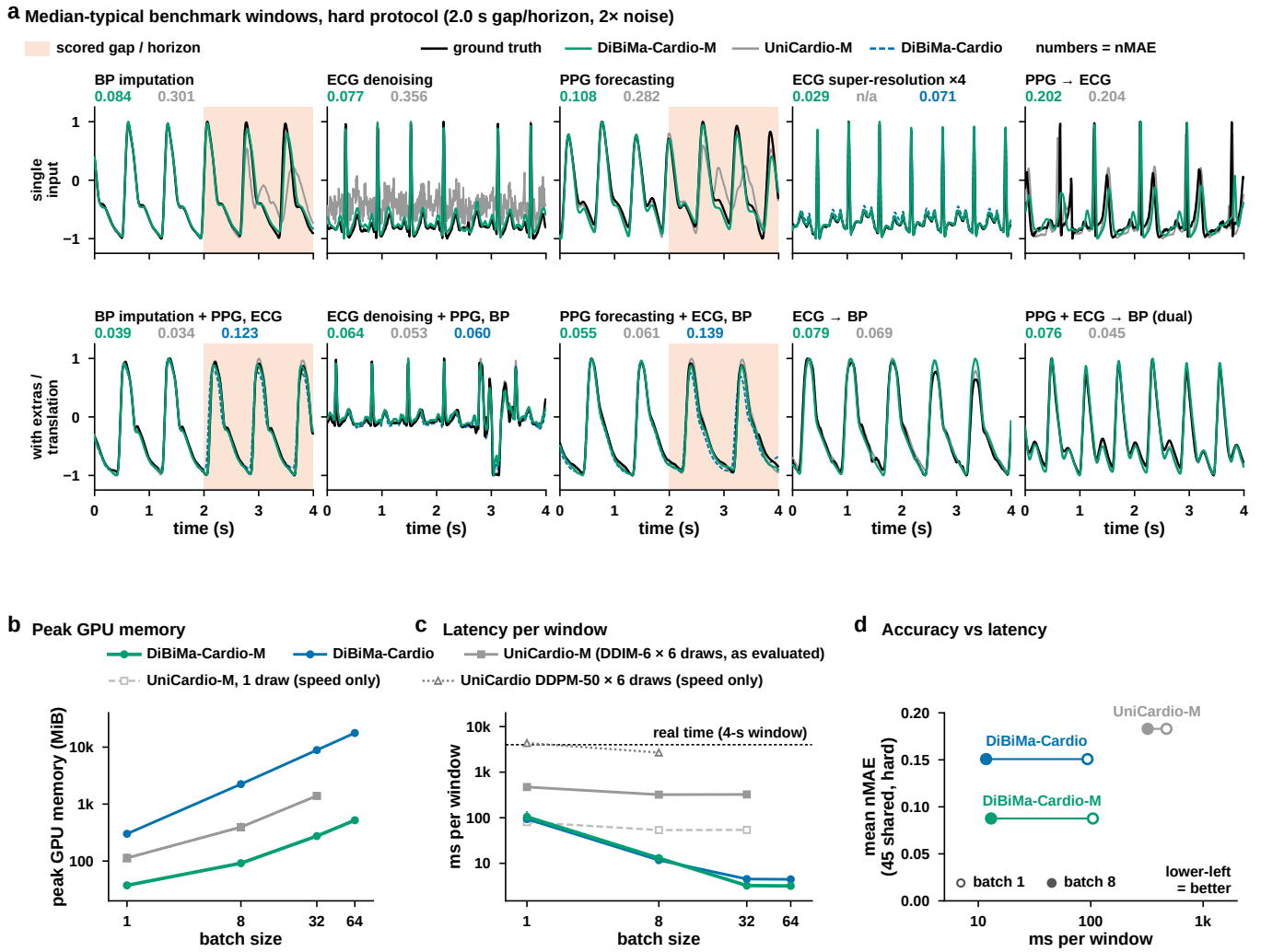


Fig. 5 Representative reconstructions and computational efficiency. Detailed panel descriptions and statistical definitions are provided in Figure legends.

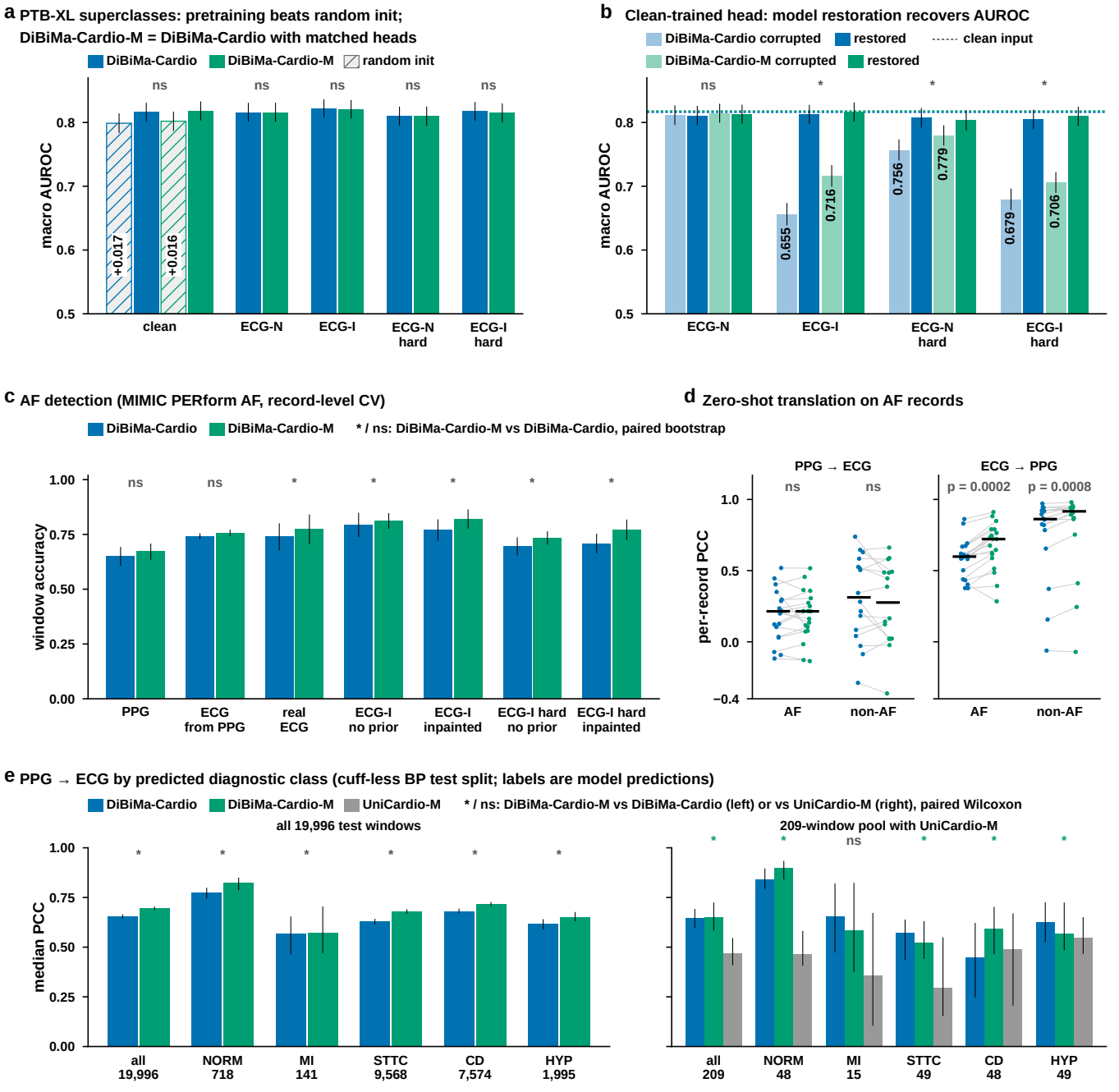


Fig. 6 Downstream classification and cross-modal fidelity. Detailed panel descriptions and statistical definitions are provided in Figure legends.