COHORT-LEVEL CONTROL FOR CONDITIONAL ECG GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional electrocardiogram (ECG) generation allows synthetic cohorts to be specified by diagnosis and patient attributes. However, diagnostic utility alone does not establish whether the ECGs reflect the requested patients. The evaluated baselines respond weakly to requested age and sex, while standard guidance for partly recorded attributes can transfer properties of the recorded population. We develop a conditional flow model that generates ECGs in an autoencoder's latent space and explicitly represents whether each attribute is supplied. A hybrid sampler contrasts supplied values for height and weight, avoiding the missing-state contrast without retraining. On PTB-XL and MIMIC-IV-ECG, the cohorts achieve train-on-synthetic, test-on-real macro AUROC of 0.903 and 0.910, respectively, and respond to age, sex, height, weight and heart rate. A controlled missingness experiment links the unintended shift to signal-dependent recording. On PTB-XL, hybrid guidance removes the observed 4.2-year age shift after a height request without reducing downstream utility.

1 INTRODUCTION

An electrocardiogram (ECG) records the heart's electrical activity. Conditional generators can supply ECGs for classifier development and evaluation in patient subgroups that are difficult to assemble from restricted clinical data. These applications require control over diagnosis and patient attributes. For example, a generated cohort of older women with a given diagnosis should also reflect the requested age and sex. Evaluating only diagnostic accuracy leaves this part of the task untested.

We apply the same measurements to generated and real cohorts to assess attribute response. This reveals two problems. The evaluated text-conditioned baselines respond weakly to requested age and sex, although they respond to heart rate. Supplying a partly recorded attribute can also change other properties of the generated ECG. In PTB-XL, requesting a population-mean height for patients whose height was unrecorded increases predicted ECG age by 4.2 years under standard guidance, despite no change to the requested age (Figure 1).

The second problem arises from the patients contributing to the "attribute absent" condition during training. They include patients whose measurement was never recorded. If recording depends on ECG features beyond the supplied conditions, these patients can differ systematically from those with a recorded measurement. Guidance between a requested value and the absent state can therefore contrast the two populations. We separate this presence contrast from the value contrast. For height and weight, our sampler guides between the requested and reference values, retaining conventional guidance for diagnosis, age, sex and heart rate within the same trained generator.

Our contributions are:

1. A conditional latent-flow generator accepting diagnoses and five patient attributes. Its cohorts achieve higher diagnostic utility than the evaluated ECG generators on PTB-XL and MIMIC-IV-ECG, at lower sampling cost than the adapted text-conditioned baselines (Section 5.2), with response to age, sex, height, weight and heart rate (Section 5.3).
2. An analysis of how a partly recorded attribute introduces a population contrast into standard guidance. We derive the absent-state mixture and test the mechanism by varying how strongly recording depends on an unconditioned ECG feature (Sections 4.2 and 5.4).

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.

3. A cohort-level evaluation that applies the same frozen measurement models to generated and real ECGs. It identifies an unintended 4.2-year age shift after a height request and shows that hybrid guidance removes the observed shift while preserving utility for fully specified cohorts (Sections 3 and 5.4).

2 Related work

Existing ECG generators support label conditioning (Alcaraz & Strodthoff, 2023), text and patient conditioning (Chung et al., 2023; Lai et al., 2025a; Wang et al., 2026), and personalised generation (Lai et al., 2025b). We evaluate how the generated ECGs respond to changes in each requested patient attribute. Our generator combines latent modelling (Rombach et al., 2022), transformer backbones (Peebles & Xie, 2023; Ma et al., 2024) and flow matching (Liu et al., 2023; Lipman et al., 2023).

Classifier-free guidance contrasts a condition with an unconditional prediction (Ho & Salimans, 2022). Composable and factor-specific variants separate multiple conditions (Liu et al., 2022; Brooks et al., 2023). The closest operation to our height and weight guidance is the contrast between minimally different prompts used by Wu & De la Torre (2024) to control an image factor while limiting changes to others. Related approaches use a contrastive objective to suppress unwanted concepts (Chang et al., 2024), or an independent condition to approximate the unconditional prediction (Sadat et al., 2025). Our sampler holds the other conditions fixed and contrasts two supplied attribute values. We use this established form of guidance to address a specific consequence of incomplete clinical records: the absent condition mixes patients with different recording histories, and its contribution can shift an unrequested ECG property.

Missingness can carry information about the underlying observation (Rubin, 1976; Agniel et al., 2018), and missing-indicator analyses have documented the resulting bias in prediction (Groenwold et al., 2012; Van Ness et al., 2023). In our generator, the same absent state represents both an unrecorded measurement and a recorded value hidden during training. The distribution learned for that state therefore depends on the recording process as well as the dropout rate. Our analysis follows this dependence into the guidance direction and tests its effect on generated signals. We supplement train-on-synthetic, test-on-real (TSTR) utility (Esteban et al., 2017) with attribute response and population measurements, applying the same instruments to real and synthetic ECGs.

3 Problem and evaluation

Let x denote an ECG and let $c = (y, a)$ contain diagnosis indicators y and patient attributes $a = (a_1, \ldots, a_K)$. The indicator R_k records whether attribute a_k is available. We represent an unavailable attribute by a distinct absent state $\varnothing_k$ so that the model can distinguish it from a supplied numerical value. We construct a synthetic cohort by sampling one or more ECGs for each request in a set $\{c_i\}$. The aim is to match the real population specified by both the diagnoses and patient attributes in these requests.

We use frozen ECG predictors g_k for age, sex, height, weight and diagnosis, and a signal-based detector for heart rate. Each instrument is applied to generated ECGs and to real ECGs selected by the same request policy. For an attribute sweep, the response slope is the slope of $g_k(\hat{x})$ against requested a_k, divided by the slope obtained when $g_k(x)$ is regressed on recorded a_k in real ECGs. A ratio of one denotes equal response slopes on generated and real ECGs. Equal slopes can still occur when generated measurements have a constant offset from the real reference. They can also conceal differences in variation between patients with the same requested attribute. We therefore compare the distributions of these measurements within each cohort as well as their response to changing requests.

We measure diagnostic utility by the macro area under the receiver operating characteristic curve (AUROC) of a classifier trained on generated ECGs and tested on held-out real ECGs. To evaluate an unintended shift, we compare the same requests before and after an attribute is supplied, using the corresponding real ECGs as a reference where available.

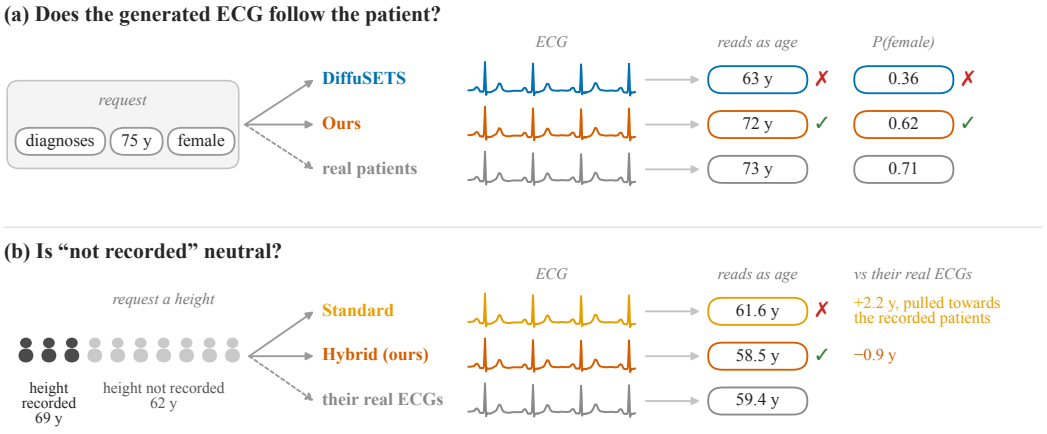


Figure 1: Two cohort-level failures on PTB-XL. (a) For requests at age 75 and female sex, frozen ECG predictors assign the adapted DiffuSETS samples lower age and female probability than real ECGs, while ours approaches the real reference. (b) Height recording selects an older training population. When a population-mean height is added for patients without a recorded height, standard guidance raises predicted age relative to those patients' real ECGs. Hybrid guidance does not. ECG traces are schematic, while numbers are measured cohort means.

4 METHOD

4.1 GENERATOR

We first train an autoencoder to map each 12-lead ECG to a compact latent representation and reconstruct the signal from it. We then train a 7.5-million-parameter diffusion transformer with adaptive layer normalisation on the encoded ECGs. At each training step, we interpolate between an ECG latent and Gaussian noise at a randomly sampled time. The transformer receives this interpolated latent, the time and the conditions, and learns the noise-minus-latent velocity by minimising squared error. To generate an ECG, we start from Gaussian noise, integrate the learned velocity backwards from noise to data, and decode the resulting latent into a 12-lead signal.

The conditioning vector contains 71 or 86 diagnostic labels, depending on the dataset, five patient attributes and their presence indicators. A presence indicator distinguishes an unavailable measurement from a supplied value at the normalisation origin. During training, we independently hide attribute factors and sometimes hide the entire condition vector. This trains the same network to predict velocities with different subsets of information. We refer to these evaluations as branches; each uses the same model parameters with a different conditioning vector.

Write v_0 for the velocity without conditions, v_Y for labels alone and v_F for the full request, all evaluated at the same latent and time. Standard factored guidance is

$$v_{\text{std}} = v_0 + w_Y(v_Y - v_0) + w_A(v_F - v_Y), \tag{1}$$

where w_Y and w_A control the label and attribute directions. The first difference measures the change in velocity when diagnoses are supplied; the second measures the additional change when patient attributes are supplied alongside those diagnoses. Weighting them separately allows the sampler to adjust the strength of each contribution. It requires three network evaluations per sampling step.

4.2 WHY STANDARD GUIDANCE LEAKS

The absent branch receives two groups of training examples: patients whose attribute was unrecorded and patients whose recorded value was hidden by dropout. Random dropout does not remove differences between these populations. Let D_k indicate that a recorded value survives dropout and set $S_k = R_k D_k$. For the other supplied conditions c_{-k}, random dropout gives the following

training distribution for the absent branch:

$$p(x \mid S_k = 0, c_{-k}) = \pi_0 p(x \mid R_k = 0, c_{-k}) + (1 - \pi_0) p(x \mid R_k = 1, c_{-k}), \tag{2}$$

where $\pi_0 = P(R_k = 0 \mid S_k = 0, c_{-k})$. The second component integrates over recorded values of a_k. Equation equation 2 follows when dropout is independent of x given recording status and c_{-k}. If R_k depends on ECG information beyond c_{-k}, the two components differ, and the absent branch represents a mixture of these recording-specific populations. If $R_k \perp x \mid c_{-k}$, the recording-specific difference vanishes, although correlations between a requested value and other attributes may remain.

Hiding a recorded height during training removes the measurement from the conditioning vector, but the ECG still comes from the population in which height was recorded. Increasing dropout changes the mixture weight without changing the ECG distributions within the two groups. Random dropout therefore preserves their differences when recording depends on signal features beyond the supplied conditions.

The guidance direction makes the unwanted term explicit. For a present request a_k and a present reference value a_k^0, evaluated with all other conditions fixed, the identity

$$v(a_k) - v(\varnothing_k) = [v(a_k) - v(a_k^0)] + [v(a_k^0) - v(\varnothing_k)] \tag{3}$$

separates a value contrast from a request-independent presence contrast. The presence term can amplify a recording-associated population shift when the attribute is supplied. We test the effect on generated ECGs by varying the recording mechanism in Section 5.4.

4.3 Measuring the presence contrast

We use a presence probe to test whether the learned contrast distinguishes patients with and without a recorded attribute. On held-out real ECG latents, we compute the norm of $v(a_k^0) - v(\varnothing_k)$ at matched noise and time, average it over repeated draws and use its negative as a recording-status score. Probe AUROC above 0.5 shows that the learned contrast contains information about recording. The probe measures association with recording status; we test for changes in unrequested attributes by sampling ECGs before and after supplying a value.

4.4 Hybrid guidance

We retain standard guidance for age, sex and heart rate, but use a value-to-value contrast for the partly recorded height and weight. Let v_A denote the request with height and weight absent and let v_R contain the same request with those attributes set to fixed training-population reference values wherever they are requested. The hybrid velocity is

$$v_{\text{hyb}} = v_0 + w_Y(v_Y - v_0) + w_A(v_A - v_Y) + w_X(v_F - v_R). \tag{4}$$

Both velocities in the final difference receive the same diagnoses, age, sex and heart rate, and the same presence indicators. Only the supplied height and weight values change. Thus, this term excludes the transition from absent to present height or weight that occurs in standard guidance. The reference values are 165 cm and 70 kg on PTB-XL, and 167 cm and 81 kg on MIMIC-IV-ECG. The sampler requires no retraining and evaluates five branches per step rather than three. When the requested values equal the references, the final difference is zero by construction. The response at other values depends on the learned value contrast, which we evaluate through attribute sweeps.

Consider a PTB-XL request for height 175 cm, with weight left absent. The two velocities in the height contrast receive 175 cm and 165 cm, respectively, with the same diagnosis, age, sex and heart rate. Height is marked present in both evaluations. At a request of 165 cm, the two inputs coincide and their difference vanishes. The mean-height experiment alone therefore cannot establish height control; the sweep tests response away from the reference.

5 Experiments

5.1 Setup

We evaluate on PTB-XL (Wagner et al., 2020; 2022) using folds 1–8 for training, fold 9 for validation and fold 10 for testing. MIMIC-IV-ECG (Gow et al., 2023) contributes 789,481 checksum-

verified ECGs. We use patient-disjoint folds and demographics from MIMIC-IV-ECG-Ext-ICD (Strodthoff et al., 2024a;b), 86 machine statements and height and weight linked from MIMIC-IV (Johnson et al., 2024; 2023). Both datasets are distributed through PhysioNet (Pollard et al., 2026). We report means and standard deviations over three independent runs for the principal comparisons. The same request lists and held-out real test sets are used for each method within a dataset.

The principal PTB-XL utility comparison requests one new ECG for each training record, reusing its diagnosis labels and available attributes. This is the "copy" protocol in the tables: the conditioning tuples are copied, and each waveform is generated from noise. The principal MIMIC-IV-ECG comparison uses 50,000 requests. We also vary the number of generated ECGs to measure how downstream utility changes with cohort size.

We adapt DiffuSETS and SE-Diff to share our generator's latent decoder and evaluation protocol. Both accept the available age, sex and heart-rate inputs; the substitutions are described in Section C. We evaluate SSSD-ECG with official weights where feasible and use the authors' reported copy-protocol TSTR, marked in Table 1. The frozen attribute predictors and heart-rate detector are calibrated on real ECGs and are never trained on generated outputs.

5.2 Utility and cost

Classifiers trained on cohorts sampled with hybrid guidance achieve TSTR macro AUROC of 0.903 on PTB-XL and 0.910 on MIMIC-IV-ECG, compared with 0.811 and 0.829 for the stronger adapted text-conditioned baseline, DiffuSETS (Table 1). The real-data references are 0.932 and 0.951. Rare-label AUROC follows the same ordering. The PTB-XL SSSD-ECG copy-protocol value of 0.840 was reported by the authors and was not measured in our classifier runs. Results at other sample budgets are given in Tables 4 and 5. The difference between our standard and hybrid samplers is small relative to their improvement over the evaluated baselines.

Hybrid guidance requires 52.5 GFLOPs per ECG at 50 steps, compared with 170.5 GFLOPs for either adapted text-conditioned baseline. Standard guidance uses 31.8 GFLOPs. Reducing the hybrid to 10 steps lowers this cost to 11.1 GFLOPs while retaining approximately the same PTB-XL TSTR as the 50-step version (Figure 2 and Table 8). On an NVIDIA GB10 at batch size 256, generating 1,000 ECGs takes 25 seconds with the 50-step hybrid, compared with 220 and 266 seconds for adapted DiffuSETS and SE-Diff. We ran the official SSSD-ECG implementation without its optional CUDA extension, which substantially increases its measured latency.

Training our model without demographic conditions gives PTB-XL TSTR 0.904, close to 0.901 with demographic conditions under standard guidance (Section E). The classifier achieves similar diagnostic performance even when the generator has no demographic inputs. We therefore measure response to these attributes separately.

5.3 Response to patient attributes

We vary each requested attribute and measure the generated ECG response (Figure 3). The adapted baselines respond to heart rate but show little response to age or sex. On PTB-XL, their age-response ratios are 0.02 and 0.03 and their sex-response ratios are 0.13; neither accepts height or weight. Our hybrid sampler responds to all five attributes, with different strengths relative to real ECGs (Table 2). Age and height responses remain weaker, at ratios of 0.73 and 0.79. Sex is close to the real response at 1.08, while weight and heart rate exceed it at 1.16 and 1.14. MIMIC-IV-ECG shows the same broad pattern: all five attributes affect the generated signals, but the strength of the response differs from the real reference, particularly for weight (Figure 4).

We measure heart rate with a signal-based detector, independently of the learned demographic predictors. On PTB-XL, mean absolute error between the requested and detected heart rate is 3.2 beats per minute for our generator, 6.8 for adapted DiffuSETS and 8.2 for adapted SE-Diff. On MIMIC-IV-ECG, the values are 2.5, 5.2 and 5.1. Applying the same detector to real ECGs gives errors of 1.9 and 2.3 beats per minute, respectively (Table 2). Published heart-rate errors use different detectors and protocols, preventing a direct comparison with these values.

Table 1: Synthetic-cohort utility. TSTR macro AUROC is measured on 71 PTB-XL statements with one ECG per training record ($n = 17{,}418$) and on 86 MIMIC-IV-ECG statements with $n = 50{,}000$ requests. Rare labels have training prevalence below 2%. Rows other than SSSD-ECG share request lists, classifier and test set. Values show mean ± SD where runs are available. *Adapted in the shared latent space. †AUROC reported by Alcaraz & Strodthoff (2023); sampling cost measured in our implementation. Dashes indicate unavailable or inapplicable entries.

	TSTR AUROC ↑				GFLOPs
	PTB-XL	rare	MIMIC-IV-ECG	rare	per sample ↓
Real data	.932 ± .006	.931 ± .007	.951 ± .001	.948 ± .001	–
Ours (hybrid)	**.903 ± .006**	**.900 ± .009**	**.910 ± .003**	**.908 ± .002**	53
Ours (standard)	.901 ± .002	.896 ± .004	**.910 ± .001**	.907 ± .001	**32**
DiffuSETS*	.811 ± .011	.808 ± .013	.829 ± .012	.826 ± .012	170
SE-Diff*	.781 ± .004	.776 ± .006	.804 ± .002	.804 ± .003	170
SSSD-ECG (published)	.840†	–	–	–	16,710

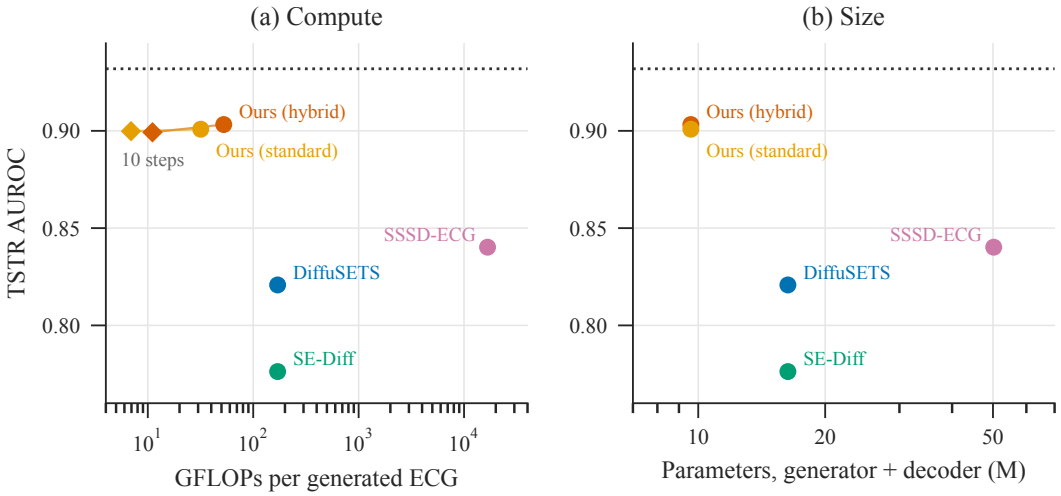


Figure 2: PTB-XL copy-protocol TSTR against sampling compute and model size. Circles show 50-step samplers, diamonds show our 10-step settings and the horizontal dotted line is the real-data classifier reference. Baseline points are adapted implementations except for SSSD-ECG's reported TSTR. FLOPs include guidance passes and decoding; Table 8 gives the measurement protocol.

Under identical request lists, the adapted baselines produce narrower or shifted attribute distributions compared with real ECGs (Figures 5 and 7). Our cohorts more closely match the age and sex distributions, with remaining discrepancies in height and weight (Section 6).

5.4 Leakage from partly recorded attributes

For PTB-XL height, the presence probe attains AUROC 0.85–0.88. To test whether supplying height changes an unrequested attribute, we add a population-mean height to 600 test requests for patients whose height was unrecorded. Predicted ECG age increases by 4.22 ± 0.81 years under standard guidance and changes by -0.18 ± 0.10 years under hybrid guidance (Table 3). Patients with a recorded height are older in the training data, consistent with the direction of the standard-guidance shift. We use MIMIC-IV-ECG as a negative control: its height probe is near chance at about 0.53, and the corresponding standard-guidance age shift has magnitude below 0.3 years. Height is partly recorded in both datasets, but the recording signal and associated age shift differ.

To test the role of the recording mechanism, we start with the fully observed PTB-XL age field and hide values at rates that differ above and below the median ECG root-mean-square (RMS)

Table 2: Attribute control under matched test requests. Slope ratios divide the generated-ECG response by the response measured with the same instrument on real ECGs. Heart-rate mean absolute error (MAE) uses a lead-II peak detector over 2,000 requests, with the detector's error on matched real ECGs given in each dataset heading. A dash denotes an attribute that the method cannot accept.

	slope ratio					HR MAE
	Age	Sex	Height	Weight	HR	(bpm)
PTB-XL (real ECGs: HR MAE 1.9)						
Ours (hybrid)	0.73	1.08	0.79	1.16	1.14	**3.2**
Ours (standard)	0.72	1.09	0.32	0.72	1.16	**3.2**
DiffuSETS*	0.02	0.13	–	–	0.74	6.8
SE-Diff*	0.03	0.13	–	–	0.58	8.2
MIMIC-IV-ECG (real ECGs: HR MAE 2.3)						
Ours (hybrid)	0.67	0.77	0.61	1.67	1.09	2.5
Ours (standard)	0.75	0.86	0.38	1.17	1.10	**2.4**
DiffuSETS*	0.01	0.13	–	–	0.86	5.2
SE-Diff*	−0.02	−0.01	–	–	0.90	5.1

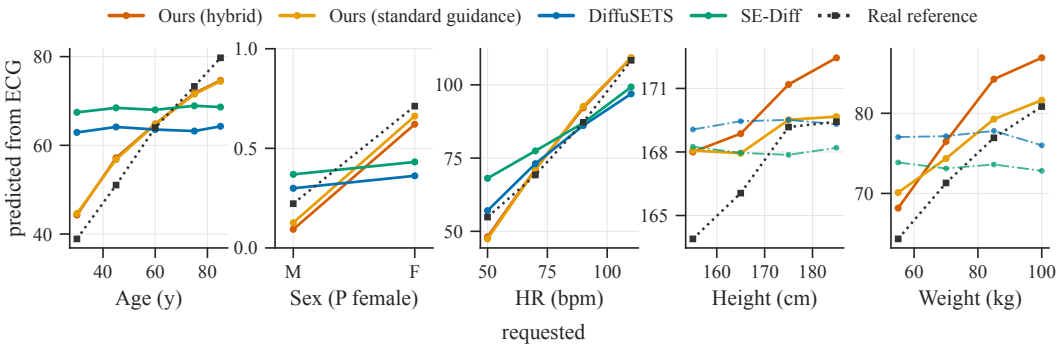


Figure 3: PTB-XL attribute response. Each panel plots the requested value against a frozen predictor or signal measurement on the generated ECG. The dotted black curve applies the same measurement to real ECGs with the corresponding recorded attribute. Dash-dot lines indicate attributes unavailable to a baseline.

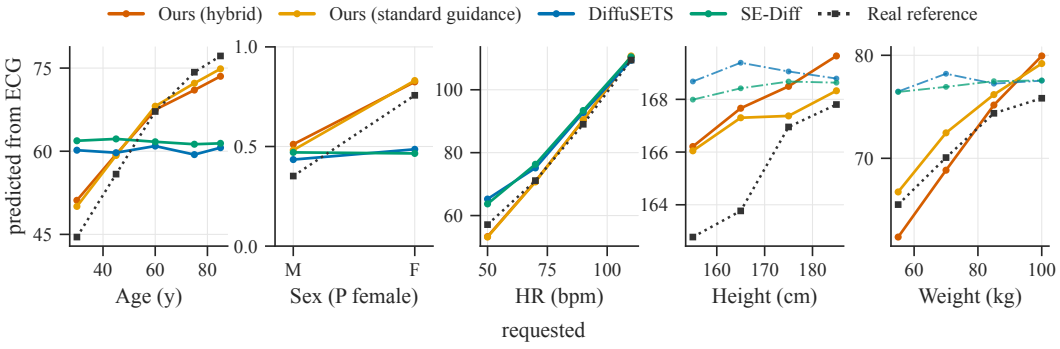


Figure 4: MIMIC-IV-ECG response curves. Requested attributes are compared with measurements on generated ECGs and the corresponding real-ECG reference, following Figure 3.

amplitude. Since amplitude is not supplied as a condition, this creates recording dependence on an unconditioned signal feature.

As the hiding-rate gap increases from 0 to 60 percentage points, requesting age changes generated amplitude by approximately 0, −6, −11, −21 and −31%, respectively (Figure 6a). The presence probe rises from 0.49 to 0.60 over the same progression. Random hiding and hiding by sex, which

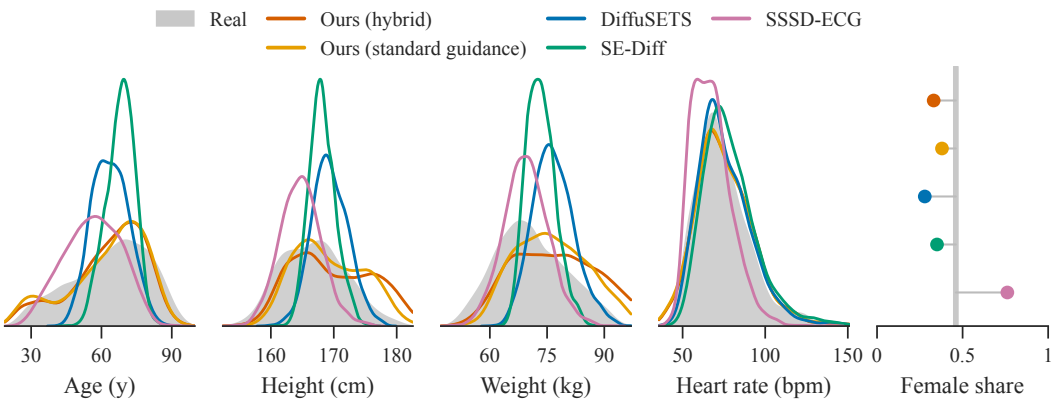


Figure 5: PTB-XL cohort distributions under the same requests. Curves show measurements on generated ECGs; grey shows measurements on real ECGs. The last panel gives female share rather than a density. Distributions include attributes unavailable as conditioning inputs to a baseline.

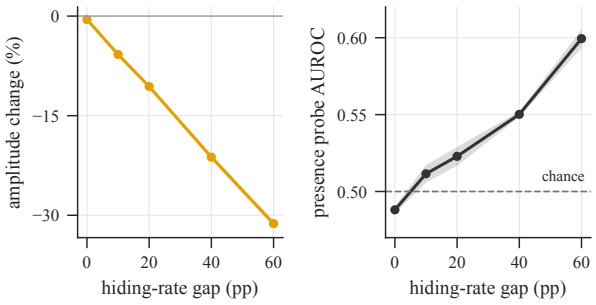
Table 3: PTB-XL alternatives to hybrid guidance. Leak is the change in ECG-predicted age after a mean-height request for patients whose height is unrecorded. TSTR uses 8,000 requests, first with recorded attributes and then with height and weight specified throughout. Height and sex columns give response slopes relative to real ECGs. Values show mean ± SD across repeated runs.

	Leak (y)	TSTR	TSTR, specified	Height	Sex
Standard guidance	4.22 ± 0.81	$.889 \pm .004$	$.873 \pm .002$	0.31 ± 0.04	1.14 ± 0.04
Hybrid (ours)	-0.18 ± 0.10	$.887 \pm .002$	$.889 \pm .004$	0.69 ± 0.10	1.10 ± 0.01
Impute before training	0.44 ± 0.22	$.888 \pm .004$	$.883 \pm .007$	0.54 ± 0.06	0.95 ± 0.02
Complete records only	0.10 ± 0.19	$.869 \pm .003$	$.864 \pm .011$	0.54 ± 0.03	1.09 ± 0.01

is supplied to the generator, do not produce the corresponding shift. Together, these interventions identify dependence on the unconditioned signal feature as the source of the recording-associated shift in this experiment. The probe is less sensitive to the weaker effects: at gaps of 10 and 20 percentage points it remains near chance despite a measurable amplitude shift.

We further compare the generated signals with the real ECGs from the same 600 PTB-XL patients without recorded height. Mean predicted age of the real ECGs is 59.4 years. After a mean-height request, mean predicted ECG age is 61.6 years under standard guidance and 58.5 under the hybrid. The generated RMS amplitudes are 0.250 and 0.212 mV, respectively, compared with 0.206 mV for the real ECGs. The difference between samplers is therefore also present in RMS amplitude, a signal measurement independent of the learned age predictor. The patients' true heights are unknown, so the real ECGs provide a reference for non-target properties but cannot establish whether the requested height distribution is correct.

At the reference value, the hybrid contrast in Equation (4) cancels. To test response away from this value, we sweep the re-



(a) Amplitude shift (b) Presence probe

Figure 6: PTB-XL recording dependence. The hiding-rate gap compares age-hiding rates above and below median ECG amplitude; zero denotes random hiding. (a) Amplitude change after requesting mean age under standard guidance. (b) Probe AUROC. Bands show variation across three runs.

quested height. On PTB-XL, the height response ratio increases from 0.31 with standard guidance to 0.69 with hybrid guidance (Table 3). We then supply height and weight throughout each 8,000-sample cohort to assess the effect on diagnostic utility. Standard-guidance TSTR falls from 0.889 to 0.873, while hybrid TSTR changes from 0.887 to 0.889 (Table 3). Imputation and complete-record training also reduce the age shift, but imputation gives weaker height response and complete-record training loses diagnostic utility. Both alternatives change the training data, whereas hybrid guidance changes sampling.

6 LIMITATIONS

Hybrid guidance improves height response and removes the observed mean-height age shift, but these gains do not extend to every distributional measure. The PTB-XL Fréchet distance in frozen classifier features increases from 15.6 to 20.3, indicating a larger discrepancy in this feature space despite the similar downstream utility. On MIMIC-IV-ECG, the weight-response slope ratio is 1.67: changes in requested weight produce larger changes in predicted weight than those measured in real ECGs. Removing the presence contrast therefore leaves errors in the learned value response. Evaluating five branches per step rather than three also increases sampling cost by a factor of 1.67.

Our comparisons with DiffuSETS and SE-Diff apply to the adapted implementations. In particular, the original DiffuSETS text interface would require an external text service incompatible with the MIMIC-IV-ECG data use agreement, so we use a local encoder. The SE-Diff adaptation uses median real beats in place of its simulator prior and omits experience retrieval. Under our common detector, adapted DiffuSETS has heart-rate MAE of 6.8 beats per minute on PTB-XL and 5.2 on MIMIC-IV-ECG; its original paper reports 6.73 and 4.15, respectively (Lai et al., 2025a). Adapted SE-Diff has MIMIC-IV-ECG error of 5.1, while its paper reports 8.43 for SE-Diff and 13.29 for DiffuSETS (Wang et al., 2026). Differences in detectors and protocols prevent a direct ranking from these figures, and heart-rate accuracy alone does not establish that the adaptations preserve diagnostic performance.

The presence probe misses weaker controlled leaks (Figure 6b). A near-chance probe score can therefore coexist with a measurable shift during sampling, as it does in the amplitude experiment. Its sensitivity limits how confidently it can rule out a recording-associated effect in another attribute or dataset.

The learned attribute predictors can also be biased, especially for height. Applying each predictor to both real and generated ECGs accounts for its response on real data, but does not establish that it has the same errors on generated signals. Our response ratios consequently describe agreement under these measurements. The independent heart-rate and RMS measurements cover two signal properties; the demographic results still depend on the learned predictors.

We also test requests for younger patients with diagnoses whose examples in this age range were excluded from generator training. Both samplers leave gaps between generated and real age or diagnosis measurements in these held-out groups (Table 7). This limits their use for constructing cohorts with diagnosis–age combinations missing from the training data. Separately, MIMIC-IV-ECG utility plateaus near 0.91 beyond 50,000 synthetic examples while the corresponding real-data classifier continues to improve. Increasing the number of generated ECGs alone does not close this utility gap over the tested sample budgets.

7 CONCLUSION

We investigated whether conditional ECG generators produce cohorts that reflect both the requested diagnoses and patient attributes. Our latent-flow generator improves diagnostic utility and measured attribute response on two datasets. The missingness analysis shows how standard guidance can introduce a recording-associated population contrast and change an unrequested attribute. For the evaluated height requests in patients whose height was unrecorded, contrasting two present values removes the observed age shift. Cohort utility is preserved when height and weight are supplied throughout.

REPRODUCIBILITY STATEMENT

The dataset definitions, patient-disjoint splits, preprocessing rules, conditioning representation, sampler settings and baseline substitutions are described in the Appendix. Main tables use the same request lists and real test sets within each dataset. PTB-XL and the MIMIC resources have separate access conditions through PhysioNet. The MIMIC-IV-ECG generation pipeline must be run in an environment authorised for the linked records. We report model and classifier variation separately where available.

ETHICS STATEMENT

The generated ECGs are research data, not clinical measurements or a basis for patient care. Diagnostic and demographic predictors used here are measurement instruments and can be biased in clinical subgroups. Strong TSTR performance does not establish clinical validity or a privacy guarantee. Users should assess memorisation and subgroup errors for their intended application and comply with the access conditions of the source data.

AI USE STATEMENT

AI tools assisted with literature discovery, research ideation, experiment code, generation of synthetic datasets, manuscript drafting and language editing. The authors selected the research questions, reviewed the analyses and remain responsible for the reported results, citations and text.

REFERENCES

Denis Agniel, Isaac S. Kohane, and Griffin M. Weber. Biases in electronic health record data due to processes within the healthcare system: Retrospective observational study. *BMJ*, 2018. doi: 10.1136/bmj.k1479.

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based conditional ECG generation with structured state space models. *Computers in Biology and Medicine*, 163:107115, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jinho Chang, Changsun Lee, Hyungjin Chung, and Jong Chul Ye. ContrastiveCFG: Guiding diffusion sampling by contrasting positive and negative concepts. *arXiv preprint arXiv:2411.17077*, 2024.

Hyunseung Chung, Jiho Kim, Joon-Myoung Kwon, Ki-Hyun Jeon, Min Sung Lee, and Edward Choi. Text-to-ECG: 12-lead electrocardiogram synthesis conditioned on clinical text reports. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv preprint arXiv:1706.02633*, 2017.

Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Jonathan W Waks, Parastou Eslami, Tanner Carbonati, Ashish Chaudhari, Elizabeth Herbst, Dana Moukheiber, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *PhysioNet*, September 2023. doi: 10.13026/4nqg-sb35. URL `https://doi.org/10.13026/4nqg-sb35`. Version 1.0.

Rolf H. H. Groenwold, Ian R. White, A. Rogier T. Donders, James R. Carpenter, Douglas G. Altman, and Karel G. M. Moons. Missing covariate data in clinical research: When and when not to use the missing-indicator method for analysis. *Canadian Medical Association Journal*, 2012. doi: 10.1503/cmaj.110977.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, 2024. doi: 10.13026/kpb9-mt58. URL https://doi.org/10.13026/kpb9-mt58. Version 3.1.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023. doi: 10.1038/s41597-022-01899-x.

Yongfan Lai, Jiabo Chen, Qinghao Zhao, Deyun Zhang, Yue Wang, Shijia Geng, Hongyan Li, and Shenda Hong. DiffuSETS: 12-lead ECG generation conditioned on clinical text reports and patient-specific information. *Patterns*, 2025a.

Yongfan Lai, Bo Liu, Xinyan Guan, Qinghao Zhao, Hongyan Li, and Shenda Hong. ECGTwin: Personalized ECG generation using controllable diffusion model. *arXiv preprint arXiv:2508.02720*, 2025b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, August 2026. doi: 10.1038/s44360-026-00096-z. URL https://doi.org/10.1038/s44360-026-00096-z.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann M. Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. In *International Conference on Learning Representations*, 2025.

Nils Strodthoff, Juan Miguel Lopez Alcaraz, and Wilhelm Haverkamp. MIMIC-IV-ECG-Ext-ICD: Diagnostic labels for MIMIC-IV-ECG. *PhysioNet*, August 2024a. doi: 10.13026/ypt5-9d58. URL https://doi.org/10.13026/ypt5-9d58. Version 1.0.1.

Nils Strodthoff, Juan Miguel Lopez Alcaraz, and Wilhelm Haverkamp. Prospects for artificial intelligence-enhanced ECG as a unified screening tool for cardiac and non-cardiac conditions: An explorative study in emergency care. *European Heart Journal – Digital Health*, 2024b. doi: 10.1093/ehjdh/ztae039.

Mike Van Ness, Tomas M. Bosschieter, Roberto Halpin-Gregorio, and Madeleine Udell. The missing indicator method: From low to high dimensions. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *PhysioNet*, November 2022. doi: 10.13026/kfzx-aw45. URL `https://doi.org/10.13026/kfzx-aw45`. Version 1.0.3.

Xiaoda Wang, Kaiqiao Han, Yuhao Xu, Xiao Luo, Yizhou Sun, Wei Wang, and Carl Yang. SE-Diff: Simulator and experience enhanced diffusion model for comprehensive ECG generation. In *International Conference on Learning Representations*, 2026. arXiv:2511.09895.

Chen Wu and Fernando De la Torre. Contrastive prompts improve disentanglement in text-to-image diffusion models. *arXiv preprint arXiv:2402.13490*, 2024.

A Data and preprocessing

PTB-XL ECGs are sampled at 100 Hz and represented with 71 diagnostic statements. MIMIC-IV-ECG records are checked against release hashes, reordered to the same lead convention and downsampled from 500 to 100 Hz. We use 86 machine statements above the selected prevalence threshold. Age, sex and patient-disjoint folds come from MIMIC-IV-ECG-Ext-ICD. Height and weight are linked to MIMIC-IV observations nearest the ECG within 365 days. These two attributes remain absent when no qualifying measurement is found. All generated and real comparison sets use the same request tuples within each experiment.

B Training, sampling and the hybrid algorithm

We train the autoencoder before the conditional flow model. The flow model uses a width-256, depth-6 transformer with adaptive layer normalisation and per-factor condition dropout of 0.15. We integrate the flow with 50 Euler steps in the principal experiments. A separate presence bit for each numerical attribute distinguishes an unavailable value from a recorded value at the normalisation origin. We use $w_Y = 2$, $w_A = 3$ and $w_X = 5$ for the reported hybrid, with fixed reference values given in Section 4.4. At each step, we evaluate all branch velocities on the same current latent and time, then form the weighted differences in Equation (4). Ten-step results use the same guidance weights with a shorter integration grid. Architecture and conditioning variants are given in Table 6.

C Baseline re-implementations

We train DiffuSETS and SE-Diff with a shared latent representation and the request interface provided by their released generation code. Both receive raw age, sex and heart rate where supported. The SE-Diff adaptation replaces the simulator prior with median real beats and does not reproduce experience retrieval. We use the paper's simulator-loss weights rather than the released-code weights, choosing the normalisation on validation data. On PTB-XL, the copy-protocol TSTR is 0.776 with the paper weights, 0.473 with released-code weights and 0.782 when the simulator terms are removed.

The text encoder available for PTB-XL cannot be used on restricted MIMIC-IV-ECG, so we substitute a local encoder. We evaluate SSSD-ECG with official weights at the 2,000-sample budget and cite its copy-protocol result from the original work. Published heart-rate errors differ across detectors and protocols, including between reports of the same method. We apply the same heart-rate detector to all baselines (Table 2).

D PREDICTORS

We obtain age, sex, height, weight and diagnostic measurements from models trained on real ECGs and frozen before evaluating generation. We apply each predictor to generated ECGs and to the corresponding real cohort. We measure heart rate from lead II with the same peak detector for every method and report its real-ECG error in Table 2. For the missingness intervention, we compute RMS amplitude directly from the signal without a learned predictor. The presence probe operates on model velocities at real ECG latents and is evaluated on held-out recording indicators.

E ADDITIONAL RESULTS

Tables 4 and 5 report the effect of sample budget alongside the principal copy and 50,000-request comparisons. Table 2 gives all measured response ratios, and Table 6 shows that the presence-associated age shift persists across the tested backbones and conditioning paths. These ablations use one run per setting, while the principal comparisons aggregate repeated training runs. In Table 7, patients aged 20–39 with each listed diagnosis are excluded from generator training, and we then request ECGs at their recorded ages. The real and generated ECGs are evaluated with the same age and diagnosis predictors, so the table's age column reports predicted ECG age even for the real patients.

On PTB-XL, training without demographics gives copy-protocol TSTR 0.904 ± 0.003, compared with 0.901 ± 0.002 for standard guidance with demographics. A label-only request to the latter changes TSTR by -0.004 ± 0.004, while assigning mismatched demographics reduces it by 0.017 ± 0.005. The cost measurements in Table 8 count the complete sampling run, including guidance branches and latent decoding.

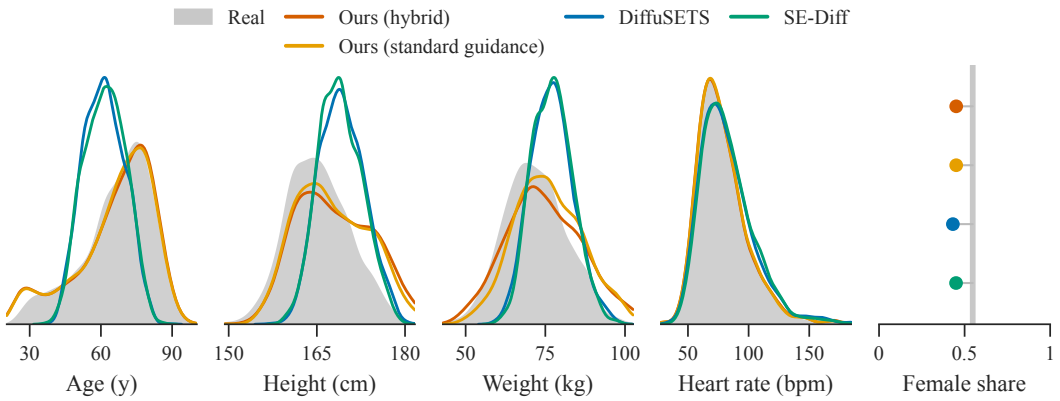


Figure 7: MIMIC-IV-ECG attribute distributions under matched request lists, measured with the same frozen instruments as Figure 5. Grey denotes real ECGs.

Table 4: PTB-XL TSTR macro AUROC on fold 10 over 71 statements. Copy uses one generated ECG per training record ($n = 17{,}418$), and rare includes 37 labels below 2% training prevalence. Other than the reported SSSD-ECG copy value, rows use matched requests, classifier and test set. The SSSD-ECG 2,000-sample row uses official weights. Values show mean ± SD where available. *Adapted in our latent space.

	$n = 2{,}000$	$n = 8{,}000$	copy	copy, rare
Real data	.858	.913	.932 ± .006	.931 ± .007
Ours (hybrid)	.850 ± .009	.887 ± .002	**.903 ± .006**	**.900 ± .009**
Ours (standard)	**.852 ± .002**	**.889 ± .004**	.901 ± .002	.896 ± .004
DiffuSETS*	.758	.790	.811 ± .011	.808 ± .013
SE-Diff*	.697	.758	.781 ± .004	.776 ± .006
SSSD-ECG	.782	–	.840	–

Table 5: MIMIC-IV-ECG TSTR macro AUROC over 86 machine statements, evaluated on 10,000 real test ECGs. Rare includes the 59 statements below 2% prevalence. All rows share requests, classifier and test set. Values show mean ± SD across repeated runs. *Adapted with a shared latent space and local text encoder.

	$n = 20{,}000$	rare	$n = 50{,}000$	rare
Real data	.930 ± .000	.926 ± .000	.951 ± .001	.948 ± .001
Ours (hybrid)	**.898 ± .002**	**.893 ± .003**	**.910 ± .003**	**.908 ± .002**
Ours (standard)	.896 ± .001	.891 ± .002	**.910 ± .001**	.907 ± .001
DiffuSETS*	.792 ± .002	.783 ± .004	.829 ± .012	.826 ± .012
SE-Diff*	.781 ± .001	.775 ± .002	.804 ± .002	.804 ± .003

Table 6: PTB-XL architecture and conditioning ablations, each trained once. The full recipe is a width-256, depth-6 flow transformer with adaptive layer normalisation, presence indicators and factor dropout of 0.15. TSTR uses 8,000 requests. Attribute columns are response slopes relative to real ECGs. Leak is the age-prediction shift after requesting mean height for an unrecorded patient. Models without presence indicators use joint guidance.

	TSTR	rare	Age	Sex	Height	Weight	Leak (y)
Ours (full recipe)	.894	.889	0.70	0.99	0.34	0.66	3.37
no presence flags, no factor dropout	.875	.865	0.63	1.07	0.11	0.63	–
presence flags, no factor dropout	.878	.867	0.66	1.13	0.23	0.79	2.43
+ independence reweighting	.892	.888	0.72	1.09	0.34	0.80	4.15
factor dropout .30	.891	.886	0.64	0.97	0.35	0.62	3.39
factor dropout .50	.885	.879	0.64	0.82	0.55	0.63	3.73
width 384, depth 8	.887	.880	0.70	1.14	0.38	0.70	2.30
width 512, depth 12	.887	.877	0.72	1.20	0.40	0.79	2.96
conditions as tokens	.881	.872	0.60	0.89	0.15	0.51	2.05
conditions by cross-attention	.877	.864	0.65	0.88	0.16	0.42	1.88
tokens + adaLN	.888	.880	0.58	0.84	0.13	0.56	2.34
DDPM backbone	.882	.874	0.64	1.01	0.27	0.62	4.26
flow-matching U-Net backbone	.894	.896	0.69	1.24	0.30	0.74	3.33

Table 7: Held-out groups of younger patients on PTB-XL. Patients aged 20–39 years with the listed diagnosis are excluded from generator training, then requested at their recorded ages. Age and diagnosis probability (P) are predicted from both real and generated ECGs; the age column does not report the patients' recorded ages. FD is Fréchet distance to the corresponding real held-out group, where lower is better. Values average repeated runs.

	Real ECGs		Ours (standard)			Ours (hybrid)		
Diagnosis	Age	P	Age	P	FD	Age	P	FD
LVH	48.3	0.24	54.1	0.34	57	54.4	0.36	58
LAFB	49.0	0.62	52.2	0.56	48	53.0	0.60	54
ASMI	53.9	0.89	57.5	0.85	69	57.8	0.89	67
PVC	52.7	0.99	54.2	0.56	91	54.0	0.58	89
1AVB	53.0	0.66	54.7	0.18	78	55.2	0.20	74

Table 8: Sampling cost on an NVIDIA GB10 at batch size 256. FLOPs and wall-clock time include all guidance passes and decoding. Parameters list generator and decoder separately. SSSD-ECG is measured without its optional CUDA extension, and its FLOP count includes analytically counted fast Fourier transforms.

	Params (M)	GFLOPs	NFE	s / 1,000
Ours (hybrid), 50 steps	7.5 + 2.1	52.5	250	25
Ours (hybrid), 10 steps	7.5 + 2.1	11.1	50	5
Ours (standard), 50 steps	7.5 + 2.1	31.8	150	14
Ours (standard), 10 steps	7.5 + 2.1	6.9	30	3
DiffuSETS*	14.2 + 2.1	170.5	1000	220
SE-Diff*	14.2 + 2.1	170.5	1000	266
SSSD-ECG	50.2	16,710.4	200	24,808