# When Simulator Targets Suppress Diversity in Conditional Diffusion: Evidence from ECG Generation

**Anonymous authors**
Paper under double-blind review

## Abstract

Matching a diffusion model's clean-signal estimate to a deterministic class template penalises within-class variation. We derive this penalty and show that the population-optimal estimate's variance contracts by $(1 + \lambda\mu_i)^{-2}$ along each eigendirection of the coupling operator. Template accuracy does not enter this factor. An exact Gaussian reverse-chain recursion predicts sampled variance ratios on synthetic data without fitted parameters. On patient-disjoint PTB-XL beats, point coupling reduces diversity to $0.66$ of real data and increases raw-signal $\mathrm{MMD}^2$ by $3.7\times$ under a shared kernel. Improving the simulator, using an oracle class mean, or drawing an independent random target does not recover the lost diversity. Training the morphology and inter-lead terms separately shows that each suppresses diversity on its own: the inter-lead term contracts the limb leads it references without improving their amplitude-normalized consistency, and a constructed beat satisfies the released inter-lead penalty exactly while violating Einthoven's law among its generated leads. The released paired evaluation ranks the most concentrated generator best. Moment matching, spread matching, and exact lead identities reduce the distributional penalty; spread matching with exact identities reaches $1.3\times$ the unregularised model's raw-signal $\mathrm{MMD}^2$. No tested coupling improves within-PTB-XL downstream classification beyond seed variability. Experiments on MIMIC-IV-ECG, multi-beat windows, and a small latent model reproduce the contraction. These results support avoiding deterministic template matching when within-class variation matters. They do not establish that simulator priors or sample-dependent constraints are ineffective in general.

## 1 Introduction

Scientific generative models must preserve variation within a condition as well as produce plausible individual samples. In ECG synthesis, variation in conduction, amplitude, and morphology affects class coverage and downstream classification. Conditional diffusion models learn these distributions from data (Ho et al., 2020; Song et al., 2021; Dhariwal & Nichol, 2021); physiological simulators offer additional structure. The McSharry model represents cardiac morphology and rhythm with a compact system of differential equations (McSharry et al., 2003). Simulator-based supervision has been used in ECG GANs (Golany et al., 2020) and text-conditioned latent diffusion (Wang et al., 2026).

The coupling matters. An exact physical identity permits many valid signals, whereas a fixed class template penalises every departure from one reference. We study the training objective

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{DDPM}} + \lambda\,\mathbb{E}\|A\hat{x}_0 - A\tau_y\|^2, \tag{1}$$

where $\hat{x}_0$ is the predicted clean signal, $\tau_y$ is fixed within class $y$, and $A$ is a linear read-out. At fixed $y$ and $t$,

$$\mathbb{E}\|A\hat{x}_0 - A\tau_y\|^2 = \underbrace{\|A\mathbb{E}[\hat{x}_0] - A\tau_y\|^2}_{\text{target mismatch}} + \underbrace{\operatorname{tr}\big(A\,\mathrm{Cov}(\hat{x}_0)A^\top\big)}_{\text{within-class variance penalty}}. \tag{2}$$

Improving the template changes the first term. The second penalises variation in every direction observed by $A$, including variation present in real data. This bias–variance identity is standard



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.

(Mathieu et al., 2016). Our contribution is to derive its consequences for diffusion denoisers and reverse sampling, then test them under controlled changes to the target and coupling.

We use phase-aligned PTB-XL beats (Wagner et al., 2020) to isolate the simulator-loss construction used in SE-Diff; the experiment does not reproduce SE-Diff's text conditioning, retrieval or latent pipeline (§5.6), and Appendices O and K examine rhythm-preserving windows and a small latent model. We evaluate distributional distances, train-synthetic/test-real (TSTR) classification (Esteban et al., 2017), learned-space precision and recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019), and clinical measurements. The study establishes four results: (I) Deterministic point coupling suppresses diversity on PTB-XL and MIMIC-IV-ECG. A more accurate simulator, an oracle class mean, and an independent random target do not remove the effect. (II) Trained separately, the morphology and inter-lead terms each suppress diversity, and their combination is worse than either with a sub-additive interaction on diversity. The template-based inter-lead term does not improve amplitude-normalized lead consistency; SE-Diff's released loss returns $5.5 \times 10^{-15}$ on a constructed beat whose generated leads violate Einthoven's law by $10.8$, and its released paired-error evaluation favours the most concentrated samples. (III) The population-optimal denoiser contracts in the eigendirections of $A^\top A$. A Gaussian covariance recursion predicts the sampled variance ratios on synthetic data without fitted parameters, but underestimates the loss on real ECGs by about $2\times$. (IV) Changing the coupling recovers most distributional fidelity, but downstream gains depend on the classifier. None of the tested couplings improves within-PTB-XL TSTR beyond seed variability; the nearest-member target matches the baseline while doubling its raw-signal distance. For the tested setting, exact lead identities preserve valid relationships without the diversity cost of a deterministic template. Our recommendation is limited to this coupling family and experimental budget. Small point-coupling advantages on cross-database classification, multi-beat classification, and the rarest class remain uncertain.

## 2 Related work

**Physics and simulator constraints.** Physics-informed learning penalises governing-equation residuals (Raissi et al., 2019; Yang et al., 2020). Diffusion methods incorporate these residuals during training (Bastek et al., 2025; Shu et al., 2023) or sampling (Chung et al., 2023). Simulator-supervised ECG models use outputs of a cardiac model (Golany et al., 2020; 2021; Wang et al., 2026); PPG-to-ECG translation can use paired targets with region-dependent weights (Shome et al., 2024). Theorem 1 applies when the target is deterministic given the condition. It does not establish collapse for sample-dependent targets or imply that suppressing violations of an exact physical relation is harmful. For a paired regression target, the relevant conditional variation may instead be measurement noise that the task intends to remove.

**ECG generation.** SE-Diff combines simulator-based supervision, latent diffusion, and retrieved clinical context (Wang et al., 2026). Conditional 12-lead diffusion models also operate without simulator targets (Lopez Alcaraz & Strodthoff, 2023). We isolate the deterministic-template loss rather than compare complete systems. Classical QRS wave-shape models (Sörnmo et al., 1981; Laguna et al., 1996) motivate the target-quality controls in Appendix E; RDDM (Shome et al., 2024) motivates the exploratory regional schedule in Appendix L.

**Squared losses and evaluation.** Squared-error regression favours conditional averages (Mathieu et al., 2016). Classifier-free guidance (Ho & Salimans, 2022), progressive distillation (Salimans & Ho, 2022), and consistency training (Song & Dhariwal, 2024) use state- or trajectory-dependent targets, so our fixed-target result does not directly apply. Kernel distances (Gretton et al., 2012), Fréchet distances (Heusel et al., 2017), precision and recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019) and TSTR (Esteban et al., 2017; Strodthoff et al., 2021) measure different properties; we report them together, calibrated against real-versus-real comparisons, since paired distance to a reference beat alone favours concentration (Appendix M).

## 3 Mechanism

### 3.1 Setting

The DDPM forward process is $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\,\varepsilon$ (Ho et al., 2020). Its noise predictor defines $\hat{x}_0 = (x_t - \sqrt{1-\bar{\alpha}_t}\,\varepsilon_\theta)/\sqrt{\bar{\alpha}_t}$. In our ECG experiments, $A = D$ is the temporal first-difference

operator and $\tau_y = T_y$ is a per-class, per-lead template. The soft inter-lead term is

$$\mathcal{L}_{\text{inter}} = \mathbb{E} \sum_{r=1}^{6} \|D\hat{x}_0^{(c_r)} - \Lambda_r(DT_y)\|^2.$$

Its reference leads come from the template, not the generated signal. We use equal nominal weights, $\lambda = 1$, as in the released code, but multiply each sample's coupling by $\bar{\alpha}_t$ (§5.6).

This weighting controls noise amplification. Writing $\delta_t = \varepsilon - \varepsilon_\theta$ gives $\hat{x}_0 - x_0 = \sqrt{(1-\bar{\alpha}_t)/\bar{\alpha}_t}\,\delta_t$. Without weighting, non-vanishing denoiser error produces large clean-estimate errors at high noise (Lemma 1). In our implementation, the unweighted coupling exceeds the diffusion loss by three orders of magnitude and training diverges; $\bar{\alpha}_t$ weighting keeps both losses $O(1)$.

### 3.2 Variance contraction

**Theorem 1.** *Fix $y$ and $t$, assume finite second moments and $\lambda \geq 0$, and optimise over measurable clean-signal estimators $\hat{g}(x_t, t, y)$. Consider $\mathcal{L}(g) = \mathbb{E}\|\hat{g} - x_0\|^2 + \lambda\mathbb{E}\|A\hat{g} - A\tau_y\|^2$ for a fixed linear $A$ and a target $\tau_y$ deterministic given $y$. Then:* ***(i)*** *the decomposition* (2) *holds, so the coupling adds the within-class covariance to the loss additively;* ***(ii)*** *the pointwise minimiser is $\hat{g}^\star = M(m + \lambda A^\top A\,\tau_y)$ with $M = (I + \lambda A^\top A)^{-1}$ and $m = \mathbb{E}[x_0 \mid x_t, y]$, giving $\mathrm{Cov}(\hat{g}^\star) = M\,\mathrm{Cov}(m)M^\top$, where $\mathrm{Cov}(m)$ is the covariance of the posterior mean over $x_t$ at fixed $y$. This is the variance resolved by the data at that timestep, not the variance of the terminal sampled distribution;* ***(iii)*** *for a unit eigenvector $v_i$ of $A^\top A$ with eigenvalue $\mu_i$,*

$$\mathrm{Var}(v_i^\top \hat{g}^\star) = (1 + \lambda\mu_i)^{-2}\ \mathrm{Var}(v_i^\top m), \tag{3}$$

*and $\mathrm{tr}\,\mathrm{Cov}(\hat{g}^\star) \leq \mathrm{tr}\,\mathrm{Cov}(m)$. For $A = I$ the standard deviation scales by $1/(1+\lambda)$.*

*Proof.* (i) is the bias–variance decomposition about the deterministic point $A\tau_y$. For (ii), setting $2(\hat{g} - m) + 2\lambda A^\top A(\hat{g} - \tau_y)$ to zero and solving gives $\hat{g}^\star$; only $m$ is random. For (iii), $Mv_i = (1 + \lambda\mu_i)^{-1}v_i$ gives (3), and $\mathrm{tr}(M^2\,\mathrm{Cov}(m)) = \mathrm{tr}(\mathrm{Cov}(m)^{1/2}M^2\,\mathrm{Cov}(m)^{1/2}) \leq \mathrm{tr}\,\mathrm{Cov}(m)$ since $M^2 \preceq I$. □

Part (ii) is ridge regression of $m$ toward $\tau_y$ in the metric of $A$. In particular, with $\mu_y = \mathbb{E}[x_0 \mid y]$, $\mathbb{E}[\hat{g}^\star] - \mu_y = (I - M)(\tau_y - \mu_y)$, so a better simulator reduces bias and leaves $M$, and the variance term, untouched; a target drawn at random independently of the sample only adds a constant (Appendix H). The eigendirectional identity and the trace bound are what hold: the Loewner strengthening $\mathrm{Cov}(\hat{g}^\star) \preceq \mathrm{Cov}(m)$ is false in general (Remark 5). For an affine decoder, replace $A$ by its composition with the decoder Jacobian. A local linearisation gives the corresponding approximation for a nonlinear decoder (Appendix K).

The theorem concerns the *denoiser's estimate*, not the reverse chain's terminal distribution, which compounds the contraction; in the linear-Gaussian case the chain can be followed exactly.

**Proposition 2** (Multi-step law, linear-Gaussian case)**.** *Let $x_0 \mid y \sim \mathcal{N}(\mu_y, \Sigma)$, let the denoiser be the population optimum of the coupled objective at every timestep, so that by Theorem 1(ii) $\hat{g}_t = M_t(m_t + \lambda_t A^\top A\,\tau_y)$ with $M_t = (I + \lambda_t A^\top A)^{-1}$, and let sampling follow the DDPM update $x_{t-1} = a_t\hat{g}_t + b_t x_t + s_t z$. Then $m_t = K_t x_t + \mathrm{const}$ with $K_t = \sqrt{\bar{\alpha}_t}\,\Sigma(\bar{\alpha}_t\Sigma + (1-\bar{\alpha}_t)I)^{-1}$, and the within-class covariance of the sampled distribution is given exactly by the recursion*

$$\mathrm{Cov}_{t-1} = C_t\,\mathrm{Cov}_t\,C_t^\top + s_t^2 I, \qquad C_t = a_t M_t K_t + b_t I, \qquad \mathrm{Cov}_T = I, \tag{4}$$

*which decouples into scalar recursions per eigendirection when $A^\top A$ and $\Sigma$ commute.*

*Proof.* $K_t$ is the Gaussian posterior gain. Each reverse step is affine in $x_t$ with noise independent of $x_t$, so covariances propagate by congruence plus $s_t^2 I$; the coupling enters only through $M_t$. □

Here $\lambda_t$ is the *effective* per-step weight set by the loss reductions, $\lambda_t = d\,\lambda(1-\bar{\alpha}_t)$ for the reductions we use (Appendix L): neither the nominal $\lambda$ nor constant in $t$, strongest at high noise where $K_t$ is smallest. §5.3 checks the recursion with no free parameter.

Table 1: Coupling definitions. $h = D\hat{x}_0$, $q_y = DT_y$, and $w_t = \lambda\bar{\alpha}_t$. Moment estimates are class-specific and internally weighted by $\bar{\alpha}_t$. Soft inter-lead losses use template parents; exact identities use generated leads I and II.

| Coupling | Simulator loss | Target dependence |
|---|---|---|
| None | $0$ | No simulator penalty. |
| Point | $w_t\|h - q_y\|^2$ | Fixed class template; penalises mean error and variance. |
| Moment matching | $\|\hat{\mu}_c - \mu_{\mathrm{ref,c}}\|^2$ $+\|\hat{\sigma}_c - \sigma_{\mathrm{ref,c}}\|^2$ | Matches bank mean and spread; finite-batch residual. |
| Spread only | $\|\hat{\sigma}_c - \sigma_{\mathrm{ref,c}}\|^2$ | Matches bank spread. |
| Nearest member | $w_t\|h - q_y^{\mathrm{nn(h)}}\|^2$ | Target depends on the sample; outside Theorem 1. |
| Inter-lead term only | $0$ | No morphology loss; the soft inter-lead term of Corollary 3 alone. |

**Corollary 3** (Template-based inter-lead penalty)**.** *The target of $\mathcal{L}_{inter}$ is built from templates, hence deterministic given $y$, so Theorem 1(i) applies verbatim to each of its six residuals: it adds a further $\sum_{r=1}^{6} \mathrm{tr}\left(D\ \mathrm{Cov}(\hat{x}_0^{(c_r)})\, D^\top\right)$, the within-class covariance of the predicted dependent limb leads $\hat{x}_0^{(c_r)}$ seen through $D$, while tying each lead to a template rather than to its generated parents (§5.6).*

### 3.3 Alternative couplings

**Moment matching** replaces the point loss with $\|\mu_{\text{gen}} - \mu_{\text{ref}}\|^2 + \|\sigma_{\text{gen}} - \sigma_{\text{ref}}\|^2$. The spread term has zero derivative at $\sigma_{\text{gen}} = \sigma_{\text{ref}} > 0$; the point loss continues to penalise nonzero spread. **Spread-only matching** removes the mean term. **Exact identities** derive the four dependent limb leads from generated leads I and II, eliminating the template-based inter-lead penalty.

We apply $\bar{\alpha}_t$ weights inside the minibatch statistics because a batch mixes noise levels with widely different clean-estimate variances. Moving the weight outside the estimator leaves substantial amplification ($631$, versus $2480$ without weighting and $O(1)$ with internal weighting). Finite class-grouped batches still penalise sample-mean variability. A first-order approximation gives a residual of order $1/n_{\text{eff}}$, potentially $40$–$50\%$ of the full penalty for the rarest class at batch size $64$ (Appendix G).

## 4 Experimental setup

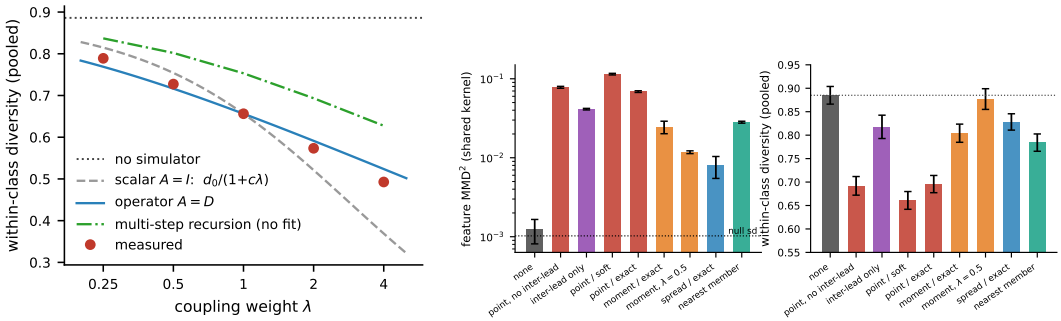
**Data and splits.** We use $4000$ PTB-XL records at $100$ Hz, five diagnostic superclasses, and $3400$ training records. Beats are R-peak aligned and resampled to $128$ cardiac-phase points. Class templates and the template bank are fitted on a separate cohort of $200$ records per class drawn from all of PTB-XL; its overlap with the generator's test sets is counted in Appendix S. Table 2 and the factorial comparison of §5.5 use patient-disjoint splits and six generator seeds (three for moment matching). The weight sweeps and the remaining controls use a record split and three seeds unless specified otherwise. Appendix B quantifies patient overlap in that split and reports the patient-disjoint replication. We also test $500$ Hz data, fewer training records, MIMIC-IV-ECG, multi-beat windows, and latent diffusion.

**Measures.** We report unbiased $\mathrm{MMD}^2$ (Gretton et al., 2012) in raw-signal space and a $108$-dimensional feature space, pooled within-class standard deviation relative to real data, and the maximum absolute lead-identity residual in mV. The raw penalty is the ratio of an arm's mean raw-signal $\mathrm{MMD}^2$ to the unregularised model's. TSTR evaluates classifiers trained on generated beats and tested on held-out real patients. Appendix S specifies the estimators and training settings.

## 5 Results

### 5.1 Coupling strength and sampled diversity

Diversity decreases monotonically over a $16\times$ range of coupling weights (Figure 1a; record split, original template cohort): from $0.886$ without coupling to $0.789$, $0.727$, $0.656$, $0.573$, and $0.493$ at




(a) Shrinkage law.

(b) Coupling geometry (patient-disjoint, ±sd).

Figure 1: **Left:** pooled within-class diversity against coupling weight (record split); both one-parameter forms are fitted at $\lambda = 1$. The scalar form over-predicts the loss at large $\lambda$; the operator form, weighted along the eigenvectors of $D^\top D$, stays within $0.03$; the parameter-free recursion (dash-dot) underestimates contraction by about $2\times$. **Right:** the patient-disjoint arms of Table 2 under its protocol (clean template cohort): shared-kernel feature $\mathrm{MMD}^2$ (log scale; dotted line, the null sd) and pooled diversity, mean±sd over seeds.

$\lambda = 0.25, 0.5, 1, 2, 4$. Two one-parameter forms are fitted at $\lambda = 1$: the scalar simplification $d_0/(1+c\lambda)$ and the operator form of Theorem 1, $d_0\sqrt{\sum_i v_i(1+k\lambda\mu_i)^{-2}/\sum_i v_i}$ with $v_i$ the real within-class variance along the $i$th eigenvector of $D^\top D$. On the four unfitted weights the operator form has mean error $0.020$ against $0.057$ for the scalar form ($0.004$ against $0.051$ under the unweighted estimator; Appendix A), still under-predicting the loss at $\lambda = 4$ by $0.031$: contraction along the operator's eigendirections rather than uniform shrinkage, with residual under-prediction at large weights. A better fit to pooled diversity supports the operator-shaped form among the two compared; it does not by itself identify directional contraction, which Appendix R probes directly.

The parameter-free recursion uses the real within-class covariance, the implemented schedule, and the effective per-lead weights, including the inter-lead term. Its mean error is $0.107$, versus $0.182$ and $0.231$ for the single-step predictions, but it underestimates the lost variance by about $2\times$. On fixed forward-noised inputs, denoiser variance ratios follow the single-step ordering with $r \geq 0.97$ at every tested timestep. Reverse-chain samples contract further (Appendix R). On real ECG both laws predict direction and ordering, not magnitude: a check of the mechanism, not a predictive law; the fixed-input probe of Appendix R uses one seed and one weight.

### 5.2 When the coupling acts

At high noise the posterior gain $K_t$ is small and little within-class variance is resolved; at low noise $K_t$ approaches the identity and coupling acts on that variation directly. The Gaussian posterior calculation resolves half the real within-class variance by $t \approx 540$ and $98\%$ by $t = 100$ (Figure 2a).

Loss reductions convert a nominal weight $\rho(t)$ into $\lambda_t = d\rho(t)/\mathrm{SNR}_t$ (Appendix L). An SNR-balanced schedule makes $\lambda_t$ constant, retaining full coupling at low noise. With point coupling and exact identities, this schedule reduces diversity to $0.58 \pm 0.03$, versus $0.68 \pm 0.02$ under $\bar{\alpha}_t$ weighting, and increases feature $\mathrm{MMD}^2$ from $0.078$ to $0.79$. A cutoff restricted to $\mathrm{SNR}_t \leq 1$ gives similar diversity ($0.67 \pm 0.04$) and lower feature $\mathrm{MMD}^2$ ($0.042 \pm 0.003$). Balancing loss magnitudes alone does not protect sampled diversity.

### 5.3 Synthetic validation of the recursion

We train a conditional DDPM on a Gaussian mixture with known anisotropic within-class covariance, a fixed class target, and a linear read-out (Appendix I). The single-step law predicts the ordering of variance ratios ($r = 0.92$–$0.94$), but overestimates every measured ratio. The reverse-chain recursion captures the difference: using the known covariance and implemented effective weights without fitting, it gives mean absolute errors of $0.005$–$0.033$ and $r \geq 0.987$.

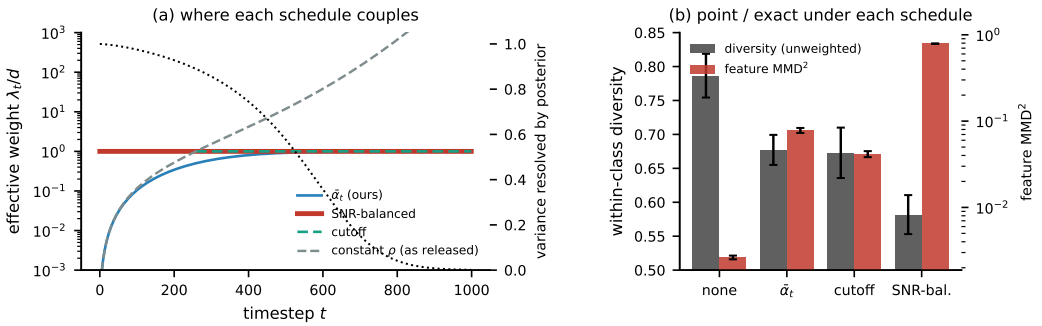



Figure 2: **(a)** Effective per-step coupling weight $\lambda_t/d$ of four schedules against the timestep, with the fraction of real within-class variance the posterior has resolved (dotted, right axis): the SNR-balanced schedule is flat and so couples at full strength where the posterior has resolved everything; the released constant-$\rho$ schedule diverges at high noise; ours and the cutoff vanish at low noise. **(b)** Point / exact under each schedule, three seeds ($\pm$sd; unweighted diversity, the estimator recorded for all four arms; feature $\mathrm{MMD}^2$ on log scale).

### 5.4 Target accuracy and coupling geometry

Gabor fits in place of Gaussian deflections reduce in-sample template RMSE by $41\%$ ($0.01768$ to $0.01042$ mV); held-out gains are $17\%$ (class-mean waveform), $15\%$ (derivative space) and $1\%$ (individual beats; Appendix E). Under point coupling, feature $\mathrm{MMD}^2$ changes from $0.1366$ to $0.1293$ while diversity stays at $0.642$–$0.643$. An oracle training-class mean also leaves diversity essentially unchanged ($0.649$ versus $0.651$), although feature $\mathrm{MMD}^2$ falls by $14\%$.

Target accuracy matters more under moment matching, whose feature $\mathrm{MMD}^2$ falls from $0.0449 \pm 0.0126$ (Gaussian) to $0.0257 \pm 0.0068$ (Gabor) and $0.0102 \pm 0.0013$ (oracle); spread-only matching has no mean target, though the bank sets its reference spread (Appendix F).

Changing the loss with a fixed template bank produces larger gains (Table 2). Point/soft to spread/exact reduces shared-kernel raw $\mathrm{MMD}^2$ from $0.0116$ to $0.0040$, a $2.9\times$ improvement, compared with $1.05\times$ from target accuracy at $\lambda = 1$ on the record split. The geometry ordering persists across a $4\times$ weight range (Appendix **??**). Enforcing exact identities on the unregularised samples alone changes neither $\mathrm{MMD}^2$ nor diversity to three decimals and reduces the RMS limb residual from $0.0027$ mV to zero.

Waveform comparisons show reduced spread with similar class means (Appendix Q). Clinical measurements identify a $7\%$ wider QRS under point coupling, a QRS KS distance of $0.31$, lower precordial R amplitudes, and a narrower frontal-axis distribution. Spread matching with exact identities largely restores these measures (Appendix P). Removing the mean term gives the best fixed-weight coupled arm, but does not improve on omitting the template.

### 5.5 Separating the two penalties

Point/soft combines two deterministic-target terms: the morphology loss, which acts on all twelve leads, and the inter-lead loss of Corollary 3, which acts on the limb leads. We train each term alone under the protocol of Table 2 and compare the four cells seed by seed; a seed fixes initialisation and training noise, so the cells at one seed are matched (Table 3).

Each term alone degrades fidelity and diversity, and the combination is worse than either. Under the shared kernel with matched draws, raw $\mathrm{MMD}^2$ rises by $0.0036 \pm 0.0008$ with the inter-lead term, $0.0056 \pm 0.0007$ with the morphology term and $0.0085$ with both. The interaction $Y_{11} - Y_{10} - Y_{01} + Y_{00}$, with its $95\%$ $t$-interval over the six matched seeds, is $-0.0035$ $[-0.0052, -0.0019]$ on the feature map and $-0.0007$ $[-0.0016, +0.0001]$ on the raw map. Pooled sd falls by $0.067 \pm 0.007$, $0.193 \pm 0.003$ and $0.224$; its interaction is $+0.036$ $[+0.027, +0.045]$: the inter-lead term removes $0.031$ of diversity when the morphology term is active against $0.067$ when it is not. The signs differ because the metrics run in opposite directions: a negative interaction on a distance and a positive

Table 2: Coupling ablation on the patient-disjoint split, one protocol for every row; templates and bank fitted on the clean cohort, which excludes every test patient (Appendix S). $\mathrm{MMD}^2$ under the shared kernel of Appendix A (one bandwidth per map fitted on real test beats, 300 beats per side, 20 evaluation draws per seed keyed by (seed, draw) so that every arm is scored against the same subsamples; null sd 0.0010 feature, 0.0009 raw), ratio of raw $\mathrm{MMD}^2$ to the unregularised model, pooled within-class sd relative to real, maximum lead-identity residual, and TSTR macro AUC of the CNN from one evaluation run over the clean-cohort samples, five classifier seeds per generator seed (none/exact from its own run; real-data reference $0.832 \pm 0.011$ from the same five classifiers as the per-class row of Appendix B). Mean$\pm$sample sd over six generator seeds ($^{\ddagger}$three); the 95% $t$-interval for a mean has half-width 1.05 sd for six seeds and 2.48 sd for three. The same arms trained on the original template cohort: Table 4. $^{\dagger}$Exact identities applied to the unregularised samples leave the distributional measures unchanged to four decimals. "Moment" denotes moment matching.

| Coupling | Leads | Feature $\mathrm{MMD}^2$ | Raw $\mathrm{MMD}^2$ | Raw ratio | Diversity | Limb (mV) | TSTR |
|---|---|---|---|---|---|---|---|
| None | Off | 0.0012 $\pm$0.0004 | 0.0031 $\pm$0.0010 | 1.0 | 0.885 | 0.072 | 0.774 $\pm$0.023 |
| None | Exact | Same$^{\dagger}$ | Same$^{\dagger}$ | 1.0 | 0.885 | 0.000 | 0.770 $\pm$0.016 |
| Point, no inter-lead term | Off | 0.0780 $\pm$0.0022 | 0.0087 $\pm$0.0017 | 2.8 | 0.692 | 0.053 | 0.750 $\pm$0.015 |
| Inter-lead term only | Soft | 0.0413 $\pm$0.0009 | 0.0068 $\pm$0.0016 | 2.2 | 0.818 | 0.074 | 0.761 $\pm$0.012 |
| Point | Soft | 0.1145 $\pm$0.0027 | 0.0116 $\pm$0.0016 | 3.7 | 0.661 | 0.057 | 0.741 $\pm$0.008 |
| Point | Exact | 0.0688 $\pm$0.0017 | 0.0086 $\pm$0.0017 | 2.7 | 0.696 | 0.000 | 0.752 $\pm$0.013 |
| Moment$^{\ddagger}$ | Exact | 0.0246 $\pm$0.0044 | 0.0046 $\pm$0.0008 | 1.5 | 0.804 | 0.000 | 0.661 $\pm$0.016 |
| Moment$^{\ddagger}$, $\lambda = 0.5$ | Exact | 0.0118 $\pm$0.0005 | 0.0032 $\pm$0.0010 | 1.0 | 0.877 | 0.000 | 0.692 $\pm$0.020 |
| Spread only | Exact | 0.0079 $\pm$0.0025 | 0.0040 $\pm$0.0011 | 1.3 | 0.828 | 0.000 | 0.724 $\pm$0.009 |
| Nearest member | Exact | 0.0282 $\pm$0.0008 | 0.0067 $\pm$0.0014 | 2.1 | 0.784 | 0.000 | 0.770 $\pm$0.013 |

interaction on diversity both mean less harm than the additive prediction. The feature-$\mathrm{MMD}^2$ and diversity intervals exclude zero and support sub-additive harm; the raw-$\mathrm{MMD}^2$ interaction is small (about one null standard deviation, 8% of the two single-term effects combined), negative in five of six seeds, and its interval includes zero, so it is not resolved; on the original template cohort it is $-0.0007$ $[-0.0014, -0.0001]$ (Appendix B). Table 3 reports the interaction per metric on its own scale; the sub-additivity on sd is consistent with the two terms constraining overlapping directions.

The pooled numbers weight each term by the leads it touches. On the four dependent limb leads the inter-lead term alone contracts sd more than the morphology term alone (0.669 against 0.706; 0.888 without coupling); on the precordial leads it barely acts (0.851 against 0.882) while the morphology term gives 0.685. The morphology term accounts for most of the pooled contraction in this implementation because it acts on twelve leads, not because it is the stronger constraint per lead. Under the implemented reductions the inter-lead term also carries about twice the per-lead effective weight of the morphology term on the limb leads ($1536/762$ against $1536/1524$ per unit nominal weight, Appendix S), so its stronger limb contraction is a property of this implementation, not of the penalty form. The absolute identity residual falls under the morphology term, from 0.072 to 0.053 mV at the maximum and from 0.0027 to 0.0023 mV in RMS, while the RMS amplitude of the dependent limb leads falls from 0.115 to 0.099 mV; the ratio is unchanged (0.0232 against 0.0230), consistent with amplitude suppression rather than improved consistency. The inter-lead term leaves the RMS residual at 0.0026 mV over an amplitude of 0.096 mV, a normalized residual of 0.0273; both terms together give 0.0252; real beats give 0.0016. The template-based inter-lead term therefore raises the normalized residual by $0.0040 \pm 0.0006$. This is the training-time counterpart of the zero-loss construction in §5.6: a penalty that references template parents can be minimised without improving consistency among generated leads. Replacing the inter-lead term by exact identities while keeping the morphology term (point / exact, trained under the same protocol) gives pooled sd 0.696 against 0.692 for the morphology-only cell, raw $\mathrm{MMD}^2$ $0.0086 \pm 0.0017$ against $0.0087 \pm 0.0017$ and

Table 3: Separating the two penalties, patient-disjoint, six matched generator seeds; the cells' overall $\mathrm{MMD}^2$, pooled diversity, maximum residual and TSTR are the corresponding rows of Table 2. Panel A adds the pooled sd ratio on the four dependent limb leads (III, aVR, aVL, aVF) and on the six precordial leads, and the residual decomposition: RMS lead-identity residual, RMS amplitude of the dependent limb leads (real beats: $0.0002$ and $0.125\,\mathrm{mV}$) and their ratio ($0.0016$ for real beats). Panel B gives the matched per-seed contrasts, mean$\pm$sd over seeds, with $\mathrm{MMD}^2$ under the shared kernel and matched draws of Table 2: the effect of each term with the other term off and on, and the interaction $Y_{11} - Y_{10} - Y_{01} + Y_{00}$; with six seeds the $95\%$ $t$-interval for a mean has half-width $1.05$ sd. The per-run $\mathrm{MMD}^2$ protocol is a sensitivity analysis in Appendix S.

**A. Lead-group diversity and the residual decomposition**

| Cell | Diversity, limb | Diversity, precordial | RMS residual (mV) | Limb RMS amplitude (mV) | normalized residual |
|---|---|---|---|---|---|
| None | 0.888 | 0.882 | 0.0027 | 0.115 | 0.0232 |
| Morphology term only | 0.706 | 0.685 | 0.0023 | 0.099 | 0.0230 |
| Inter-lead term only | 0.669 | 0.851 | 0.0026 | 0.096 | 0.0273 |
| Both (point / soft) | 0.594 | 0.675 | 0.0023 | 0.090 | 0.0252 |
| None, exact identities | 0.888 | 0.882 | 0.0000 | 0.115 | 0.0000 |
| Morphology, exact identities (point / exact) | 0.729 | 0.684 | 0.0000 | 0.101 | 0.0000 |

**B. Matched contrasts**

| Metric | Morphology, inter-lead off | Morphology, inter-lead on | Inter-lead, morphology off | Inter-lead, morphology on | Interaction |
|---|---|---|---|---|---|
| Raw $\mathrm{MMD}^2$ | $+0.0056$ $\pm0.0007$ | $+0.0048$ $\pm0.0003$ | $+0.0036$ $\pm0.0008$ | $+0.0029$ $\pm0.0003$ | $-0.0007$ $\pm0.0008$ |
| Feature $\mathrm{MMD}^2$ | $+0.0768$ $\pm0.0024$ | $+0.0732$ $\pm0.0029$ | $+0.0400$ $\pm0.0009$ | $+0.0365$ $\pm0.0012$ | $-0.0035$ $\pm0.0016$ |
| Diversity, all | $-0.193$ $\pm0.003$ | $-0.157$ $\pm0.007$ | $-0.067$ $\pm0.007$ | $-0.031$ $\pm0.004$ | $+0.036$ $\pm0.009$ |
| Diversity, limb | $-0.182$ $\pm0.006$ | $-0.075$ $\pm0.003$ | $-0.220$ $\pm0.007$ | $-0.112$ $\pm0.002$ | $+0.107$ $\pm0.007$ |
| Diversity, precordial | $-0.196$ $\pm0.003$ | $-0.177$ $\pm0.008$ | $-0.030$ $\pm0.008$ | $-0.010$ $\pm0.004$ | $+0.020$ $\pm0.009$ |
| Residual | $-0.0002$ $\pm0.0006$ | $-0.0020$ $\pm0.0002$ | $+0.0040$ $\pm0.0006$ | $+0.0022$ $\pm0.0004$ | $-0.0018$ $\pm0.0005$ |
| TSTR | $-0.024$ $\pm0.017$ | $-0.020$ $\pm0.007$ | $-0.013$ $\pm0.020$ | $-0.009$ $\pm0.011$ | $+0.004$ $\pm0.014$ |

TSTR $0.752 \pm 0.013$ against $0.750 \pm 0.015$; its feature $\mathrm{MMD}^2$ is lower ($0.0688 \pm 0.0017$ against $0.0780 \pm 0.0022$), its dependent-limb diversity higher ($0.729$ against $0.706$), and the identities hold exactly. On the distributional measures the comparison is established: deriving the limb leads leaves raw distance and pooled diversity where the morphology-only cell has them and lowers the feature distance, whereas the template-based inter-lead term worsens all three. On classification the two cells are indistinguishable at six seeds (paired difference $+0.002 \pm 0.007$), and the inter-lead term's own classification effect below is within seed variability, so that part of the comparison remains uncertain. Downstream effects are small. The morphology term lowers TSTR by $0.024 \pm 0.017$ (interval excluding zero), the inter-lead term by $0.013 \pm 0.021$ (interval including zero), with interaction $+0.004 \pm 0.014$; the combination loses $0.033$ against the unregularised model. The inter-lead term's classification effect is within seed variability, so most of the distributional damage in Table 3 is not visible to this classifier.

### 5.6 Check against the released implementation

We tested SE-Diff's released loss functions on constructed inputs without modifying them.[1] The code applies coupling to the predicted clean signal, uses no $\bar{\alpha}_t$ weighting, and gives the three losses equal weight. Our training uses its own $\bar{\alpha}_t$-weighted reduction because the unweighted version diverges under our settings.

[1] `https://github.com/ignite-abd/SE-Diff`. The audit covers the loss functions; it does not cover the complete latent, retrieval, and text pipeline.

A constructed beat gives $5.5 \times 10^{-15}$ for all six template-based inter-lead residuals while violating Einthoven's law among its generated leads by $10.8$ in dimensionless maximum first-difference units. Thus, zero loss does not imply consistency among generated leads. The released paired-error evaluation also favours the point-coupled samples in our experiment, despite their worse distributional distance and TSTR (Appendix M).

### 5.7 Utility and limits of the evidence

**Distance calibration.** The unbiased $\mathrm{MMD}^2$ statistic can be negative under equal distributions. A positive real-versus-real comparison is not a lower bound. With a shared kernel, matched counts, and $500$ null splits, the unregularised model lies $1.2$ null standard deviations from zero in feature space; point coupling lies $112$ away. We report ratios of mean raw distances without subtracting a null reference. Feature and raw distances agree on ordering but differ substantially in magnitude (Appendix A).

**Downstream utility.** TSTR trains on $1024$ generated beats and evaluates macro one-vs-rest AUC on $600$ held-out real beats. On the patient-disjoint split, the unregularised model exceeds point coupling by $0.033 \pm 0.022$ with the CNN (paired over six matched generator seeds, $t \approx 3.6$). Spread/exact improves over point/soft with the ResNet by $0.032 \pm 0.013$ (original template cohort), but not with the CNN ($-0.017 \pm 0.012$). Moment matching remains worse after tuning $\lambda$ to $0.5$. Nearest-member coupling matches the unregularised model under both classifiers while retaining a $2.1\times$ raw-distance penalty.

The feature-space Fisher ratio rises from $0.09$ (real) and $0.10$ (no coupling) to $0.52$ under point coupling, $0.13$ for spread/exact; the sharper class separation may explain why restored variation does not consistently improve classification. Class-specific results differ, particularly for HYP, with only 13 test records (Appendix B). Neither classifier utility nor distributional fidelity alone captures both properties.

**Other datasets and budgets.** Retraining on subject-disjoint MIMIC-IV-ECG with refitted templates reproduces the point penalty ($3.2\times$ raw distance, diversity $0.72$ of the baseline, $0.026$ lower AUC) while spread/exact matches the baseline on all three; its $2494$ training and $441$ test beats carry keyword-derived labels, which limits cross-dataset comparison. PTB-XL-trained generators evaluated on MIMIC give point coupling a small TSTR advantage ($0.008 \pm 0.006$) that three seeds do not establish (Appendix N). At training sizes $100$, $500$ and $1000$ with test data and gradient steps fixed, point coupling remains worse on both distances and diversity; spread/exact improves raw distance at $1000$ records but not feature distance (Appendix C).

**Representation and scope.** Phase alignment removes rhythm. Four-second multi-beat windows retain it and still show a $2.8\times$ raw-distance penalty under beat-wise template coupling, with diversity falling from $0.70$ to $0.57$ and little change in RR summaries or TSTR (Appendix O); the tests constrain within-beat templates, not a simulator loss on rhythm. The Gaussian recursion is exact under its assumptions but underestimates contraction in trained ECG models; finite-batch moment estimates, small rare-class samples and three to six generator seeds further limit the conclusions.

## 6 Conclusion

Squared matching of a diffusion model's clean estimate to a deterministic class template penalises within-class variation in directions observed by the coupling operator, regardless of template accuracy. Patient-disjoint experiments on PTB-XL and MIMIC-IV-ECG demonstrate this contraction; the PTB-XL findings persist when template fitting excludes all test patients. The factorial experiment separates morphology-induced contraction across leads from the additional limb-lead contraction caused by template-based inter-lead matching, which does not improve amplitude-normalized consistency. Exact lead identities enforce consistency without a measured distributional penalty. Directional probes support operator-dependent contraction, while the covariance recursion, exact under Gaussian assumptions, closely matches trained synthetic models but underestimates contraction on real ECGs. In the tested setting, these findings support enforcing physical identities exactly and omitting deterministic template matching. Nearest-member targets preserve classification utility but still alter the generated distribution; broader sample-dependent objectives and rhythm-level simulator constraints remain open questions.

## REFERENCES


Jan-Hendrik Bastek, WaiChing Sun, and Dennis M. Kochmann. Physics-informed diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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

Tomer Golany, Daniel Freedman, and Kira Radinsky. SimGANs: Simulator-based generative adversarial networks for ECG synthesis to improve deep ECG classification. In *International Conference on Machine Learning (ICML)*, 2020.

Tomer Golany, Daniel Freedman, and Kira Radinsky. ECG ODE-GAN: Learning ordinary differential equations of ECG dynamics via generative adversarial learning. In *AAAI Conference on Artificial Intelligence*, 2021.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Pablo Laguna, Raimon Jané, Salvador Olmos, Nitish V. Thakor, Hervé Rix, and Pere Caminal. Adaptive estimation of QRS complex wave features of ECG signal by the hermite model. *Medical and Biological Engineering and Computing*, 34:58–68, 1996.

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.

Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *International Conference on Learning Representations (ICLR)*, 2016.

Patrick E. McSharry, Gari D. Clifford, Lionel Tarassenko, and Leonard A. Smith. A dynamical model for generating synthetic electrocardiogram signals. *IEEE Transactions on Biomedical Engineering*, 50(3):289–294, 2003.

Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Debaditya Shome, Pritam Sarkar, and Ali Etemad. Region-disentangled diffusion model for high-fidelity PPG-to-ECG translation. In *AAAI Conference on Artificial Intelligence*, 2024.

Dule Shu, Zijie Li, and Amir Barati Farimani. A physics-informed diffusion model for high-fidelity flow field reconstruction. *Journal of Computational Physics*, 478:111972, 2023.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations (ICLR)*, volume 2024, pp. 15078–15097, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Leif Sörnmo, Per Ola Börjesson, Mats-Erik Nygards, and Olle Pahlm. A method for evaluation of QRS shape features using a mathematical model for the ECG. *IEEE Transactions on Biomedical Engineering*, 28(10):713–717, 1981.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2021.

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.

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 *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Liu Yang, Dongkun Zhang, and George Em Karniadakis. Physics-informed generative adversarial networks for stochastic differential equations. *SIAM Journal on Scientific Computing*, 42(1): A292–A317, 2020.

# A MEASUREMENT LIMITS

The interpretation of distributional distances depends on null calibration, feature choice, and the diversity estimator.

**Null calibration.** The per-run protocol in Appendix S estimates the bandwidth from each pooled comparison sample, uses 400 beats per side, and reports one 300/300 real-versus-real split as a reference. The unbiased $\mathrm{MMD}^2$ statistic has expectation zero under equal distributions and can be negative (Gretton et al., 2012). Subtracting a single positive reference does not provide a stable denominator for a ratio.

We therefore fix one RBF bandwidth per feature map using real test beats, use 300 beats per side for every comparison, draw 500 real-versus-real splits, and score each arm with 20 draws per seed, the draws keyed by (generator seed, draw index) so that every arm at a seed is scored against the same subsamples; Tables 2 and 3 report this protocol for every patient-disjoint arm. The feature and raw null means are $-0.0001$ and $0.0000$, with standard deviations $0.0010$ and $0.0009$; $59\%$ and $56\%$ of splits are negative. Across six patient-disjoint seeds on the clean template cohort, the unregularised model scores $0.0012$ in feature space and $0.0031$ in raw space, point coupling scores $0.114$ and $0.0116$, and spread/exact scores $0.0079$ and $0.0040$ (original cohort: $0.112$ and $0.0115$, $0.0056$ and $0.0037$). The shared-kernel raw ratios are $3.7\times$ and $1.3\times$, compared with feature ratios of $93\times$ and $6.4\times$; on the record split the raw ratios are $3.3\times$ and $1.2\times$. With unmatched draws the raw ratio moved by about $0.2$ between draw sets, one reason we report absolute values beside ratios. We report ratios of means, without subtracting the real-versus-real reference.

**Feature dependence.** The feature map contains per-lead mean, standard deviation, peak-to-peak amplitude, and six log-PSD bands. It does not fully represent waveform timing, shape, or inter-lead dependence. Standardising these 108 statistics by their real-data spread can emphasise shifts that appear smaller across the 1536 raw coordinates. For point coupling, the corresponding reported penalties are $46\times$ and $3.2\times$ (Figure 3). To assess sensitivity, we shrink real within-class deviations by $k = 0.90, 0.80, 0.70$. Feature $\mathrm{MMD}^2$ becomes $0.0032, 0.0109, 0.0258$, and raw $\mathrm{MMD}^2$ becomes $0.0069, 0.0175, 0.0347$. Both detect a $10\%$ reduction relative to the null standard deviation of $0.0010$.

**Diversity estimation.** Giving every class equal weight makes the estimate sensitive to rare classes: two halves of the real test set give ratios of $1.027$ and $0.875$, with only 13 records in the smallest class. Pooling within-class sums of squares reduces that spread ($1.047$ and $0.946$). It also agrees with an independent projection-based estimate ($0.886$ versus $0.888$). The adjustment varies across runs, from $1.024$ to $1.206$, and increases the baseline–point gap from $0.129$ to $0.197$. We use pooled diversity unless a table explicitly identifies the unweighted estimator. Differences below approximately $0.03$ are not interpreted.

**Functional-form sensitivity.** The operator form weights each eigendirection of $D^\top D$ by the real within-class variance along that eigenvector. An earlier version of the fit paired the eigenvalues with variances along principal components instead; under that mismatch the two forms tied at $0.057$ under the pooled estimator. With the weights measured along the eigenvectors (`lambda_law2.py`), the operator form has mean error $0.020$ against $0.057$ for the scalar form on the four unfitted weights under the pooled estimator, and $0.004$ against $0.051$ under the unweighted one; its errors are $-0.020$, $-0.011$, $+0.018$ and $+0.031$ at $\lambda = 0.25$, $0.5$, $2$ and $4$, so it over-predicts the loss at small weights and under-predicts it at large ones. The comparison is between two one-parameter fits pinned at $\lambda = 1$; it favours the operator reading without identifying the functional form. The mismatch was confined to this fit: the eigendirection analyses of Appendices I, K and R project onto the eigenvectors of their operators (`toy_mechanism.py`, `multistep_law.py`, `sediff_latent.py`, `denoiser_cov.py`), and no other reported number used the paired weights.

# B PATIENT-DISJOINT EVALUATION

PTB-XL has $1.155$ records per patient, and $11.2\%$ of patients have repeated recordings. In the record split, 31 of 597 test patients ($5.2\%$) also occur in training. We repeat the principal comparison by splitting on `patient_id`, verify zero overlap, and train six generator seeds per arm.

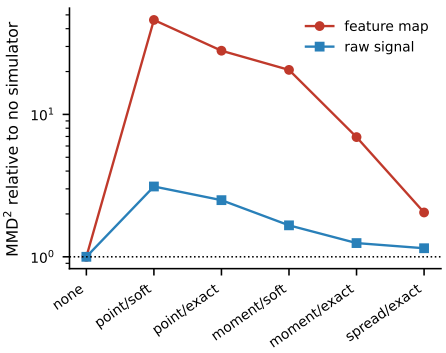



Figure 3: The two distances rank the couplings identically but disagree on magnitude by an order of magnitude. The feature map standardises 108 summary statistics; the raw map uses 1536 signal coordinates.

| | record split, three seeds | | | patient-disjoint, six seeds | | |
|---|---|---|---|---|---|---|
| coupling | raw penalty | diversity | TSTR AUC | raw penalty | diversity | TSTR AUC |
| none | 1.0 | 0.886 | 0.788 ±0.032 | 1.0 | 0.885 | 0.774 ±0.023 |
| point / soft | 3.2 | 0.656 | 0.732 ±0.021 | 3.7 | 0.661 | 0.741 ±0.008 |
| spread / exact | 1.2 | 0.831 | 0.729 ±0.015 | 1.3 | 0.828 | 0.724 ±0.009 |

The patient-disjoint columns are the clean-cohort arms of Table 2; the record split uses the original cohort. The two splits agree within 0.005 on diversity for all three arms; the raw penalties are 3.2 against 3.7 and 1.2 against 1.3. The unregularised model's TSTR advantage over point coupling decreases from approximately 0.06 to 0.033 (paired $0.033 \pm 0.022$, $t \approx 3.6$ over six matched seeds). Per-seed AUC ranges are 0.74–0.80 and 0.73–0.75. No coupling shows a reliable advantage on either split, but the magnitude of the baseline advantage remains uncertain. All patient-disjoint real-data references use training patients only.

**Per class, the macro gap is not uniform.** One-vs-rest AUC per superclass, mean over generator seeds (three on the record split, six patient-disjoint, three for moment matching), each averaged over five classifier seeds:

| split | training set | NORM | MI | STTC | CD | HYP |
|---|---|---|---|---|---|---|
| record | real, 1024 beats | 0.926 | 0.839 | 0.864 | 0.875 | 0.658 |
| | none | 0.890 | 0.787 | 0.806 | 0.802 | 0.654 |
| | point / soft | 0.839 | 0.747 | 0.709 | 0.805 | 0.560 |
| | spread / exact | 0.889 | 0.765 | 0.781 | 0.741 | 0.471 |
| patient | real, 1024 beats | 0.901 | 0.800 | 0.819 | 0.868 | 0.773 |
| | none | 0.874 | 0.787 | 0.752 | 0.784 | 0.673 |
| | point / soft | 0.791 | 0.727 | 0.700 | 0.756 | 0.733 |
| | spread / exact | 0.870 | 0.777 | 0.724 | 0.738 | 0.514 |
| | Moment/exact, $\lambda = 0.5$ | 0.855 | 0.700 | 0.616 | 0.730 | 0.557 |
| | Moment/exact, $\lambda = 1$ | 0.818 | 0.622 | 0.569 | 0.706 | 0.588 |

Point coupling trails the unregularised model by 4–10 AUC points on NORM, MI, and STTC, while matching or exceeding it on CD in the record split (0.805 versus 0.802) and HYP in the patient split (0.733 versus 0.673). Spread/exact recovers NORM and MI but performs poorly on HYP (0.47–0.51) and CD. HYP has only 13 test and 80 training records, so its class-specific ordering remains descriptive. Similar macro AUCs can conceal these opposing class effects.

**A standard diagnostic classifier separates the couplings.** The classifier above is a small three-block CNN. Repeating the six-seed patient-disjoint evaluation on the original template cohort with a standard-diagnostic 1-D ResNet (convolutional stem, three stages of two residual blocks at widths

32/64/128, global average pooling, 395k parameters; same optimiser, epochs, class weights and five classifier seeds):

| training set | macro AUC | NORM | MI | STTC | CD | HYP |
|---|---|---|---|---|---|---|
| real, 1024 beats | 0.831 ±0.018 | 0.889 | 0.814 | 0.837 | 0.849 | 0.729 |
| none | 0.778 ±0.009 | 0.867 | 0.772 | 0.750 | 0.768 | 0.732 |
| point / soft | 0.724 ±0.016 | 0.744 | 0.724 | 0.672 | 0.754 | 0.727 |
| spread / exact | 0.756 ±0.013 | 0.870 | 0.770 | 0.752 | 0.746 | 0.643 |

With the ResNet, the unregularised model exceeds point coupling by $0.054 \pm 0.011$ AUC, and spread/exact exceeds point coupling by $0.032 \pm 0.013$ (standard deviation of paired seed differences). The CNN difference for spread/exact versus point is $-0.017 \pm 0.012$ on the clean cohort (Table 2) and $-0.006 \pm 0.017$ on the original cohort. Spread/exact matches the baseline on NORM, MI, and STTC but remains worse on HYP. Thus the classification benefit of restored variation depends on the classifier.

**Augmentation.** With patient-disjoint real training data, 1024 real beats alone give 0.833 AUC; adding 1024 generated beats gives 0.828 (none), 0.820 (point), and 0.827 (spread). With 3400 real beats, the corresponding values are 0.873 alone and 0.864, 0.867, and 0.861 after augmentation. Tuned moment matching ($\lambda = 0.5$) gives 0.821 and 0.860, versus real-only references of 0.835 and 0.870 in that run. Label fidelity is 0.54, 0.55, 0.53, and 0.51 for none, point, spread, and moment matching, compared with 0.58–0.61 on real test beats.

**Template-cohort sensitivity.** Tables 2 and 3, Figure 1b and the patient-disjoint numbers of §5 use the clean template cohort, fitted after excluding every test patient (zero overlap by record and by patient; Appendix S). The same arms were first trained on the original cohort, whose overlap with the test patients is counted there; Table 4 compares the two cohorts arm by arm and seed by seed. The distributional ordering of the arms is the same on both, and raw $\mathrm{MMD}^2$ and pooled diversity change by at most one seed sd for every arm: point/soft has raw $\mathrm{MMD}^2$ 0.0116 against 0.0115 ($\Delta = +0.0001 \pm 0.0001$) and pooled diversity 0.661 on both cohorts; spread/exact has raw $\mathrm{MMD}^2$ 0.0040 against 0.0037 ($+0.0003 \pm 0.0003$) and diversity 0.828 against 0.831 ($-0.003 \pm 0.007$). Two quantities do change. Feature $\mathrm{MMD}^2$ is higher on the clean cohort by $0.0028 \pm 0.0004$ for point/soft and $0.0023 \pm 0.0027$ for spread/exact, two to three null standard deviations, consistent with the clean templates fitting the training beats slightly less well (RMSE 0.0105 against 0.0104 mV). The TSTR of spread/exact falls from 0.739 to 0.724 ($-0.014 \pm 0.010$; the six-seed interval excludes zero), so its downstream deficit to the unregularised model is larger on the clean cohort ($0.050 \pm 0.018$, paired) than on the original; no other arm's TSTR change has an interval excluding zero. Moment matching at $\lambda = 1$ moves the most at three seeds (feature $\mathrm{MMD}^2$ $-0.0049 \pm 0.0054$, diversity $+0.011 \pm 0.001$, TSTR $-0.016 \pm 0.016$). For moment matching at $\lambda = 1$, the clean-minus-original changes over three seeds are $-0.0049 \pm 0.0054$ in feature $\mathrm{MMD}^2$, $+0.011 \pm 0.001$ in diversity, and $-0.016 \pm 0.016$ in TSTR (mean $\pm$ SD). The diversity increase is statistically resolved, with an approximate 95% paired $t$-interval of $[0.0085, 0.0135]$. The corresponding intervals for feature $\mathrm{MMD}^2$ and TSTR include zero. The template-free baseline is identical in both cohorts but is scored in two evaluation runs: $0.774 \pm 0.023$ in the run over the clean-cohort samples, which Table 2 reports, and $0.770 \pm 0.017$ in the run over the original-cohort samples, which the original columns of Table 4 and the ResNet comparison above use; the difference is within one classifier-seed sd, and each table takes the baseline from its own run. The factorial contrasts of Table 3 recomputed on the original cohort (`factorial_summary.py` on `runs/psplit_seeds`) agree to the third decimal, with interactions $+0.036$ $[+0.028, +0.043]$ on sd, $-0.0044$ $[-0.0057, -0.0030]$ on feature $\mathrm{MMD}^2$ and $-0.0007$ $[-0.0014, -0.0001]$ on raw $\mathrm{MMD}^2$; the raw interval excludes zero on the original cohort and includes it on the clean one, and the exact-identity control matches the morphology-only cell on both.

Table 4: Template-cohort sensitivity. Each template-using patient-disjoint arm of Table 2 (clean cohort: templates and bank fitted after excluding every test patient, Appendix S) against the same arm at the same generator seeds trained on the original cohort. Shared-kernel $\mathrm{MMD}^2$ with the kernel and matched draws of Table 2, pooled within-class diversity and CNN TSTR; $\Delta$ is the paired per-seed difference clean minus original, mean$\pm$sd over seeds. The $95\%$ $t$-interval for a mean has half-width $1.05$ sd with six seeds and $2.48$ sd with three ($^{\ddagger}$). The template-free baseline is identical in both cohorts; its TSTR is $0.774 \pm 0.023$ in the evaluation run that scores the clean arms (Table 2) and $0.770 \pm 0.017$ in the runs that score the original arms here.

**A. Distances**

| | Feature $\mathrm{MMD}^2$ | | | Raw $\mathrm{MMD}^2$ | | |
|---|---|---|---|---|---|---|
| Coupling | original | clean | $\Delta$ | original | clean | $\Delta$ |
| Point / soft | $0.1117 \pm 0.0025$ | $0.1145 \pm 0.0027$ | $+0.0028 \pm 0.0004$ | $0.0115 \pm 0.0016$ | $0.0116 \pm 0.0016$ | $+0.0001 \pm 0.0001$ |
| Point / exact | $0.0674 \pm 0.0019$ | $0.0688 \pm 0.0017$ | $+0.0014 \pm 0.0005$ | $0.0086 \pm 0.0017$ | $0.0086 \pm 0.0017$ | $-0.0000 \pm 0.0001$ |
| Point, no inter-lead term | $0.0765 \pm 0.0021$ | $0.0780 \pm 0.0022$ | $+0.0015 \pm 0.0002$ | $0.0087 \pm 0.0016$ | $0.0087 \pm 0.0017$ | $+0.0000 \pm 0.0001$ |
| Inter-lead term only | $0.0408 \pm 0.0009$ | $0.0413 \pm 0.0009$ | $+0.0004 \pm 0.0001$ | $0.0067 \pm 0.0015$ | $0.0068 \pm 0.0016$ | $+0.0000 \pm 0.0001$ |
| Spread only / exact | $0.0056 \pm 0.0022$ | $0.0079 \pm 0.0025$ | $+0.0023 \pm 0.0027$ | $0.0037 \pm 0.0011$ | $0.0040 \pm 0.0011$ | $+0.0003 \pm 0.0003$ |
| Nearest member / exact | $0.0290 \pm 0.0008$ | $0.0282 \pm 0.0008$ | $-0.0008 \pm 0.0003$ | $0.0067 \pm 0.0014$ | $0.0067 \pm 0.0014$ | $-0.0000 \pm 0.0001$ |
| Moment / exact$^{\ddagger}$ | $0.0294 \pm 0.0073$ | $0.0246 \pm 0.0044$ | $-0.0049 \pm 0.0054$ | $0.0053 \pm 0.0005$ | $0.0046 \pm 0.0008$ | $-0.0007 \pm 0.0004$ |
| Moment / exact$^{\ddagger}$, $\lambda = 0.5$ | $0.0119 \pm 0.0001$ | $0.0118 \pm 0.0005$ | $-0.0001 \pm 0.0005$ | $0.0034 \pm 0.0010$ | $0.0032 \pm 0.0010$ | $-0.0002 \pm 0.0001$ |

**B. Diversity and downstream classification**

| | Pooled diversity | | | TSTR (CNN) | | |
|---|---|---|---|---|---|---|
| Coupling | original | clean | $\Delta$ | original | clean | $\Delta$ |
| Point / soft | $0.661 \pm 0.001$ | $0.661 \pm 0.001$ | $+0.000 \pm 0.002$ | $0.745 \pm 0.012$ | $0.741 \pm 0.008$ | $-0.003 \pm 0.005$ |
| Point / exact | $0.696 \pm 0.002$ | $0.696 \pm 0.002$ | $+0.000 \pm 0.002$ | $0.753 \pm 0.007$ | $0.752 \pm 0.013$ | $-0.000 \pm 0.009$ |
| Point, no inter-lead term | $0.692 \pm 0.001$ | $0.692 \pm 0.001$ | $-0.000 \pm 0.001$ | $0.753 \pm 0.014$ | $0.750 \pm 0.015$ | $-0.003 \pm 0.007$ |
| Inter-lead term only | $0.818 \pm 0.001$ | $0.818 \pm 0.001$ | $-0.000 \pm 0.001$ | $0.763 \pm 0.014$ | $0.761 \pm 0.012$ | $-0.002 \pm 0.008$ |
| Spread only / exact | $0.831 \pm 0.007$ | $0.828 \pm 0.007$ | $-0.003 \pm 0.007$ | $0.739 \pm 0.019$ | $0.724 \pm 0.009$ | $-0.014 \pm 0.010$ |
| Nearest member / exact | $0.784 \pm 0.002$ | $0.784 \pm 0.002$ | $+0.000 \pm 0.002$ | $0.772 \pm 0.013$ | $0.770 \pm 0.013$ | $-0.002 \pm 0.008$ |
| Moment / exact$^{\ddagger}$ | $0.793 \pm 0.001$ | $0.804 \pm 0.001$ | $+0.011 \pm 0.001$ | $0.677 \pm 0.006$ | $0.661 \pm 0.016$ | $-0.016 \pm 0.016$ |
| Moment / exact$^{\ddagger}$, $\lambda = 0.5$ | $0.870 \pm 0.007$ | $0.877 \pm 0.007$ | $+0.007 \pm 0.007$ | $0.701 \pm 0.020$ | $0.692 \pm 0.020$ | $-0.010 \pm 0.009$ |

## C TRAINING-SET SIZE

We vary the training-set size while holding the test set fixed. Each configuration receives approximately 15 930 gradient steps so that differences in optimisation budget do not confound the comparison.

Each cell uses three generator seeds. The templates and bank are the same 200-record-per-class cohort as in the main experiments (Appendix S), so at $n_{\text{train}} = 100$ the coupled arms' targets summarise more records than the generator sees; the comparison is between arms at fixed $n_{\text{train}}$ with that prior held fixed, not a comparison of information budgets. We compare the unregularised model, point/soft, and spread/exact at $n_{\text{train}} = 100, 500, 1000, 3400$. Ratios compare mean distances with the unregularised model at the same size; Welch statistics use variation across generator seeds.

| $n_{\text{train}}$ | coupling | feature $\text{MMD}^2$ | raw $\text{MMD}^2$ | raw ratio | $t$ (df) | diversity |
|---|---|---|---|---|---|---|
| 100 | none | 0.0101 ±0.0018 | 0.0168 ±0.0020 | 1.0 | — | 0.783 ±0.019 |
| | point / soft | 0.0243 ±0.0038 | 0.0377 ±0.0048 | 2.24 | +7.0 (2.7) | 0.642 ±0.018 |
| | spread / exact | 0.0141 ±0.0029 | 0.0199 ±0.0011 | 1.18 | +2.3 (3.1) | 0.749 ±0.010 |
| 500 | none | 0.0077 ±0.0010 | 0.0361 ±0.0033 | 1.0 | — | 0.750 ±0.010 |
| | point / soft | 0.0609 ±0.0009 | 0.0688 ±0.0024 | 1.91 | +13.8 (3.7) | 0.564 ±0.011 |
| | spread / exact | 0.0088 ±0.0015 | 0.0352 ±0.0016 | 0.97 | −0.4 (2.9) | 0.735 ±0.005 |
| 1000 | none | 0.0051 ±0.0007 | 0.0178 ±0.0011 | 1.0 | — | 0.804 ±0.019 |
| | point / soft | 0.1157 ±0.0024 | 0.0354 ±0.0028 | 1.99 | +10.2 (2.6) | 0.602 ±0.019 |
| | spread / exact | 0.0082 ±0.0014 | 0.0137 ±0.0018 | 0.77 | −3.4 (3.2) | 0.783 ±0.026 |
| 3400 | none | 0.0022 ±0.0008 | 0.0058 ±0.0009 | 1.0 | — | 0.874 ±0.021 |
| | point / soft | 0.1224 ±0.0021 | 0.0184 ±0.0011 | 3.17 | +15.6 (3.9) | 0.651 ±0.018 |
| | spread / exact | 0.0071 ±0.0019 | 0.0068 ±0.0008 | 1.18 | +1.5 (3.9) | 0.824 ±0.015 |

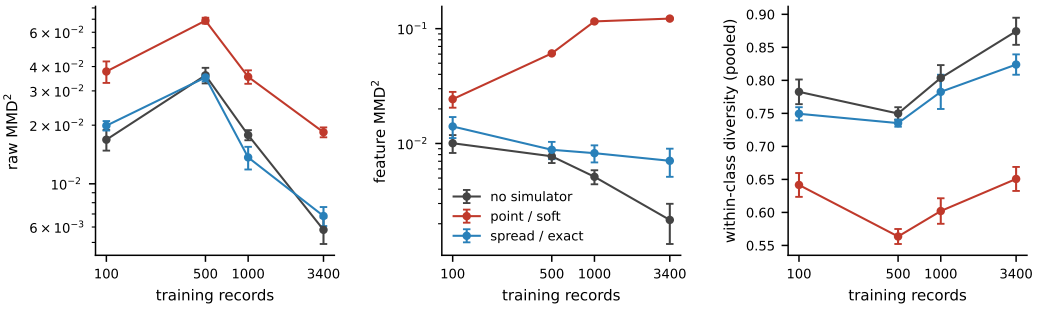



Figure 4: Low-data ablation: raw $\text{MMD}^2$, feature $\text{MMD}^2$ and pooled within-class diversity against training-set size, three generator seeds per cell (±sd), test set fixed and gradient steps matched. The point coupling is worse on both distances and on diversity at every size; the corrected coupling tracks the unregularised model on the raw distance and stays above it on the feature map.

**Point coupling.** Raw-distance penalties are 2.24, 1.91, 1.99, and 3.17 at the four training sizes, with Welch $t \geq 7$. Feature-distance penalties range from 2.4× to 56× as the baseline improves. Diversity relative to the baseline is 0.82, 0.75, 0.75, and 0.74. Reducing the training set does not remove the penalty in these experiments.

**Spread matching with exact identities.** Raw ratios are 1.18, 0.97, 0.77, and 1.18, with Welch statistics +2.3, −0.4, −3.4, and +1.5. The benefit at $n = 1000$ does not recur at neighbouring sizes, and feature distance is worse at every size (1.4, 1.1, 1.6, and 3.2× baseline). Diversity remains

0.96–0.98 of baseline. These uncertainties cover generator seeds, not finite test-set sampling. Raw distance itself is non-monotone in training size for all three arms, so the controlled comparisons are between arms at fixed size.

## D REPLICATION AT 500 HZ

Both sampling rates are resampled to 128 phase points, so the 500 Hz experiment changes source resolution while keeping the model dimensions fixed. Approximately 350 raw samples are downsampled, compared with approximately 70 samples upsampled at 100 Hz. We refit templates at 500 Hz.

| | Diversity | | raw $\mathrm{MMD}^2$ penalty | |
|---|---|---|---|---|
| coupling | 100 Hz | 500 Hz | 100 Hz | 500 Hz |
| none | **0.886** | **0.886** | 1.0 | 1.0 |
| point / soft | 0.656 | 0.632 | 3.2 | 4.5 |
| spread / exact | 0.831 | 0.814 | 1.2 | 1.4 |

Baseline diversity agrees to three decimals across rates; coupled-arm differences are at most 0.024. Raw penalties retain their ordering, although the 500 Hz values use one generator seed. TSTR is $0.776 \pm 0.007$ without coupling, $0.739 \pm 0.014$ for point coupling, and $0.736 \pm 0.024$ for spread matching, against $0.846 \pm 0.008$ with real training data. The single real-versus-real reference decreases from 0.0017 to 0.0004; one draw does not establish a change in the null distribution. Diversity outside the simulator-defined wave windows increases from 0.713 to 0.817, while the windows change little. Interpolation at 100 Hz is one possible explanation.

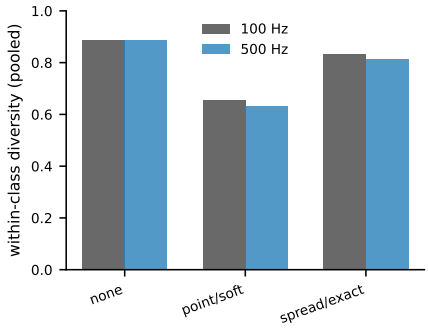



Figure 5: Within-class diversity replicates across sampling rate.

## E TEMPLATE-FITTING CONTROLS

We fit ten wave-shape families to held-out halves of 300 records per lead. The 25-parameter Gabor family and a 14-parameter split-normal–Hermite hybrid outperform Gaussian deflections under paired signed-rank tests ($p < 10^{-13}$). Gabor templates reduce in-sample class-mean RMSE by 41% (0.01768 to 0.01042 mV). Held-out improvements are 17% in class-mean waveform space, 15% in derivative space, and 1% against individual beats. These fits vary target quality while keeping the coupling fixed.

Larger weights worsen feature distance and reduce diversity for each coupled arm. Raw-distance rankings are less uniform: moment matching at $\lambda = 0.5$ gives a ratio of 0.87 and diversity 0.890, but feature distance remains $3.6\times$ baseline. The patient-disjoint three-seed replication (clean template cohort, shared kernel of Table 2) gives raw $\mathrm{MMD}^2$ $0.0032 \pm 0.0010$, versus $0.0031 \pm 0.0010$ for the six-seed baseline, and diversity $0.877 \pm 0.022$ versus 0.885. Feature $\mathrm{MMD}^2$ remains higher (0.0118 versus 0.0012), and TSTR is lower ($0.692 \pm 0.020$ versus $0.774 \pm 0.023$). At $\lambda = 1$, moment matching gives $0.661 \pm 0.016$ AUC. Tuning improves two distributional measures without producing a downstream advantage.

Table 5: Coupling-weight sensitivity on the record split with the original template cohort (generator seed 0). $\mathrm{MMD}^2$ uses an RBF bandwidth estimated separately for each comparison. Raw ratios are relative to the unregularised model. Diversity is the pooled within-class standard deviation relative to real data. For point/soft, $\lambda$ scales both morphology and inter-lead penalties; for the exact-identity arms, it scales the simulator loss.

| Coupling | $\lambda$ | Feature $\mathrm{MMD}^2$ | Raw $\mathrm{MMD}^2$ | Raw ratio | Diversity |
|---|---|---|---|---|---|
| None | — | 0.0026 | 0.0061 | 1.00 | 0.886 |
| Point / soft | 0.5 | 0.0703 | 0.0141 | 2.30 | 0.727 |
| | 1 | 0.1199 | 0.0192 | 3.13 | 0.656 |
| | 2 | 0.1739 | 0.0276 | 4.49 | 0.573 |
| Moment / exact | 0.5 | 0.0093 | 0.0053 | 0.87 | 0.890 |
| | 1 | 0.0180 | 0.0077 | 1.25 | 0.806 |
| | 2 | 0.0524 | 0.0117 | 1.90 | 0.714 |
| Spread / exact | 0.5 | 0.0050 | 0.0065 | 1.06 | 0.861 |
| | 1 | 0.0053 | 0.0071 | 1.15 | 0.831 |
| | 2 | 0.0107 | 0.0098 | 1.60 | 0.777 |

**Exact identities without a template.** Deriving dependent limb leads from the unregularised model's generated I and II changes feature distance, raw distance, and diversity by less than $0.001$ on every seed and split. The RMS limb residual (the four frontal-plane identities of Table 3) decreases from $0.0027\,\mathrm{mV}$ to zero. TSTR is $0.789 \pm 0.024$ on the record split and $0.770 \pm 0.016$ patient-disjoint, approximately $0.06$ and $0.05$ above spread/exact. These results support exact identities without template regularisation in the tested setting.

**Combined corrections.** Relative feature-distance ratios to point/soft are $0.615$ for exact identities, $0.468$ for moment matching, and $0.198$ for both. The corrections remove different penalties, but multiplying these ratios is not a justified independence model for nonlinear training.

**Sliced-Wasserstein control.** A single-seed sliced-Wasserstein coupling with exact identities gives diversity $0.803$, feature $\mathrm{MMD}^2$ $0.0426$, and a raw penalty of $12.8$. Matching marginal spread alone does not ensure fidelity. The cause of this result remains unresolved; the rank-based estimator's treatment of per-sample $\bar{\alpha}_t$ weights is one possible factor.

## F Simulator accuracy under each coupling

We compare Gaussian and Gabor template banks with an oracle based on the empirical training class mean and real training beats. All other settings remain fixed. Table 6 separates target-fit errors from generation outcomes: in-sample class-mean fit, held-out class-mean error in derivative space, and held-out per-beat error measure different sources of mismatch.

Point coupling gains about $5\%$ in feature $\mathrm{MMD}^2$ from Gabor templates, with nearly unchanged diversity. Spread-only feature distances agree within seed variability, but Gaussian-bank diversity is lower ($0.677$ versus $0.740$); its reference bank is itself $4\%$ narrower. Moment matching improves from $0.0449 \pm 0.0126$ to $0.0257 \pm 0.0068$ in feature $\mathrm{MMD}^2$ (Welch $t \approx 2.3$, three seeds), while diversity rises by $0.046$. Thus template quality affects the mean-matching term, and bank spread can affect spread-only matching even without a mean target.

The oracle has zero in-sample class-mean error by construction. Its held-out derivative-space error is $0.0115$, reflecting the train–test difference. Under moment matching, feature $\mathrm{MMD}^2$ decreases from approximately $0.045$ to $0.026$ to $0.010$ across Gaussian, Gabor, and oracle targets. Under point coupling, the oracle lowers feature $\mathrm{MMD}^2$ by $14\%$ ($0.1224$ to $0.1058$), while pooled diversity remains $0.649$–$0.651$. Removing training-mean error does not remove the covariance penalty.

Table 6: Simulator accuracy under each coupling (Moment: moment matching). Template error in mV RMSE: in-sample fit, held-out class mean in $D$-space, held-out per beat; diversity is the unweighted within-class sd ratio, the estimator available for every arm here. Three seeds where $\pm$ is given, one otherwise.

**A. Template errors (mV RMSE)**

| Template | Training fit | Held-out $D$-space | Held-out per beat |
|---|---|---|---|
| Gaussian | 0.0177 | 0.0170 | 0.0623 |
| Gabor | 0.0104 | 0.0146 | 0.0617 |
| Oracle (class mean) | 0 | 0.0115 | 0.0611 |

**B. Generation outcomes**

| Coupling | Template | Feature $\mathrm{MMD}^2$ | Diversity |
|---|---|---|---|
| Point / soft | Gaussian | 0.1366 | 0.642 |
| Point / soft | Gabor | 0.1293 | 0.643 |
| Spread / exact | Gaussian | $0.0061 \pm 0.0017$ | $0.677 \pm 0.022$ |
| Spread / exact | Gabor | $0.0075 \pm 0.0015$ | $0.740 \pm 0.018$ |
| Moment / exact | Gaussian | $0.0449 \pm 0.0126$ | $0.692 \pm 0.014$ |
| Moment / exact | Gabor | $0.0257 \pm 0.0068$ | $0.738 \pm 0.015$ |
| Point / soft | Oracle | $0.1058 \pm 0.0026$ | $0.632 \pm 0.025$ |
| Moment / exact | Oracle | $0.0102 \pm 0.0013$ | $0.680 \pm 0.029$ |

## G THE FINITE-BATCH MOMENT ESTIMATOR

Section 3.3 states the moment-matching and spread losses at the population level. The implemented estimator is finite-batch and class-grouped, and that changes what it penalises. For each class $c$ present in a minibatch of $B = 64$ beats, with $n_c$ beats of that class and reliability weights $p_i \propto \bar{\alpha}_{t_i}$ summing to one over them,

$$\hat{\mu}_c = \sum_i p_i \, Dh_i, \qquad \hat{\sigma}_c^2 = \sum_i p_i \, (Dh_i - \hat{\mu}_c)^2,$$

and the loss is $\frac{1}{|C|} \sum_c \|\hat{\mu}_c - \mu_{\text{ref},c}\|^2 + \|\hat{\sigma}_c - \sigma_{\text{ref},c}\|^2$ over the classes with $n_c \geq 2$; classes with fewer beats are skipped. Conditional on the timesteps in the batch, for beats independent given their timesteps and a fixed reference,

$$\mathbb{E}\|\hat{\mu}_c - \mu_{\text{ref}}\|^2 = \Big\| \textstyle\sum_i p_i \, \mathbb{E}[Dh_i \mid t_i] - \mu_{\text{ref}} \Big\|^2 + \textstyle\sum_i p_i^2 \, \operatorname{tr} \operatorname{Cov}(Dh_i \mid t_i),$$

which reduces to $\|\mathbb{E}Dh - \mu_{\text{ref}}\|^2 + \frac{1}{n_{\text{eff}}} \operatorname{tr} \operatorname{Cov}(Dh)$ with $n_{\text{eff}} = (\sum_i p_i^2)^{-1} \leq n_c$ only if the conditional covariance were common across timesteps. It is not: both the weights $p_i \propto \bar{\alpha}_{t_i}$ and the distribution of $Dh_i$ depend on $t_i$, and the weighting concentrates on the low-noise beats, whose conditional covariance is the largest (the posterior has resolved the within-class variance there). The exact statement is the weighted sum above; the $1/n_{\text{eff}}$ form is a first-order summary that we use only to size the effect. In the same way, the weighted population variance $\hat{\sigma}_c^2$ is biased low by a factor whose i.i.d. value is $1 - 1/n_{\text{eff}}$ (the usual correction is the multiplier $n_{\text{eff}}/(n_{\text{eff}} - 1)$), which the spread term then compensates by pushing $\sigma_{\text{gen}}$ above $\sigma_{\text{ref}}$. With the training class frequencies (0.46, 0.25, 0.18, 0.084, 0.024) the expected per-batch counts are 30, 16, 11, 5.4 and 1.5: under the first-order summary the residual penalty is 3–9% of the full variance term for the three large classes, about 19% for CD, and for HYP the class is coupled in only 44% of batches (Poisson probability of $n_c \geq 2$) and then with $n_{\text{eff}} \approx 2$–3, i.e. 40–50% of the penalty. These percentages are approximations from the batch composition, not measurements under the actual estimator, and the timestep dependence could move them in either direction. The qualitative point stands: the moment-matching arms are *not* penalty-free on the rare classes. This is consistent with, but does not explain, the corrected coupling's low AUC on HYP in Appendix B, and part of that rare-class failure may be an artefact of the estimator as implemented. The $n_{\text{eff}}/(n_{\text{eff}} - 1)$ factor corrects the weighted variance estimator when the beats share one conditional distribution; it does not make the square-root estimate of the standard deviation unbiased, it does not apply as stated to the timestep-heterogeneous mixture above,

and it leaves the $\sum_i p_i^2 \operatorname{tr}\operatorname{Cov}(Dh_i \mid t_i)$ component of the squared sample-mean objective in place. Removing that component requires a U-statistic form (products of independent beats) or subtraction of the estimated sampling variance, together with timestep-stratified statistics and a per-class minimum batch count. We did not run a batch-size sensitivity study, and state this as a limitation of the correction as implemented.

## H STOCHASTIC AND SAMPLE-DEPENDENT TARGETS

An independent stochastic target retains the same variance penalty. For $\tau$ independent of $\hat{g}$,

$$\mathbb{E}\|A\hat{g} - A\tau\|^2 = \|A(\mathbb{E}\hat{g} - \mathbb{E}\tau)\|^2 + \operatorname{tr}\big(A \operatorname{Cov}(\hat{g}) A^\top\big) + \operatorname{tr}\big(A \operatorname{Cov}(\tau) A^\top\big),$$

so the penalty on $\hat{g}$ is unchanged, the target's own spread enters as a constant, and the population optimum is that of Theorem 1 with $\tau_y$ replaced by $\mathbb{E}\tau$: the coupling regresses onto the bank mean. Theorem 1 does not cover a target *coupled to the sample*, such as the nearest bank member or a moment-matching loss, where the target moves with $\hat{g}$.

We tested the prediction by replacing the class template with a random member of the 32-template class bank, redrawn at every optimisation step, everything else fixed (three seeds).

| coupling | target | feature $\mathrm{MMD}^2$ | raw $\mathrm{MMD}^2$ | Diversity |
|---|---|---|---|---|
| point / soft | deterministic template | 0.1224 $\pm$0.0021 | 0.0184 | 0.651 $\pm$0.018 |
| point / soft | random bank member | 0.1226 $\pm$0.0054 | 0.0181 | 0.638 $\pm$0.017 |
| point / exact | deterministic template | 0.0798 $\pm$0.0048 | — | — |
| point / exact | random bank member | 0.0699 $\pm$0.0069 | 0.0137 | 0.679 $\pm$0.014 |
| none | — | 0.0022 $\pm$0.0008 | 0.0058 | 0.874 $\pm$0.021 |

The random-target and deterministic-target results agree within seed variability. With soft interlead residuals, feature $\mathrm{MMD}^2$ is $0.1226$ versus $0.1224$, and diversity is $0.638$ versus $0.651$. With exact identities, feature $\mathrm{MMD}^2$ is $0.0699$ versus $0.0798$; unweighted diversity is $0.668$ versus $0.680$. Independent target sampling does not recover the lost spread.

**Nearest-member coupling.** At each optimisation step, we match each clean estimate to the nearest of its class's 32 bank members in derivative space and enforce exact identities. The target depends on the estimate, so Theorem 1 does not apply. Across six patient-disjoint seeds on the clean template cohort, shared-kernel feature $\mathrm{MMD}^2$ is $0.0282 \pm 0.0008$, raw $\mathrm{MMD}^2$ is $0.0067 \pm 0.0014$ ($2.1\times$ baseline), and pooled diversity is $0.784 \pm 0.018$, versus $0.885$ without coupling. TSTR matches the baseline: $0.770 \pm 0.013$ versus $0.774 \pm 0.023$ for the CNN (paired $-0.004 \pm 0.014$), and $0.773 \pm 0.009$ versus $0.778 \pm 0.009$ for the ResNet on the original cohort. A sample-dependent target can retain classification utility while still altering the generated distribution.

## I SYNTHETIC-MIXTURE CHECK OF THE RECURSION

We train a small conditional DDPM ($T = 200$) on an anisotropic class-conditional Gaussian mixture, with and without a deterministic-target loss. The single-step factors correlate with sampled variance ratios at $r = 0.94$ for $\lambda = 0.5$ and $r = 0.92$ for $\lambda = 2$, but overestimate all ratios. With $A = I$, the single-step predictions are $0.444$ and $0.111$, whereas measured averages are approximately $0.24$ and $0.075$. Across the eight directions, measured ratios span $0.16$–$0.42$ and $0.04$–$0.16$ despite identical single-step factors.

Using the known within-class covariance and effective weights $\lambda_t = d\lambda(1 - \bar{\alpha}_t)$, the reverse-chain recursion gives mean absolute errors of $0.026$ and $0.005$ for $A = I$, and $0.033$ and $0.022$ for generic $A$, at the two weights. Correlations are $0.987$–$0.999$, compared with single-step errors of $0.05$–$0.21$. The largest-variance direction also deviates most from the single-step prediction because its posterior gain differs. Residual under-prediction remains empirical: network estimation error can increase or decrease covariance.

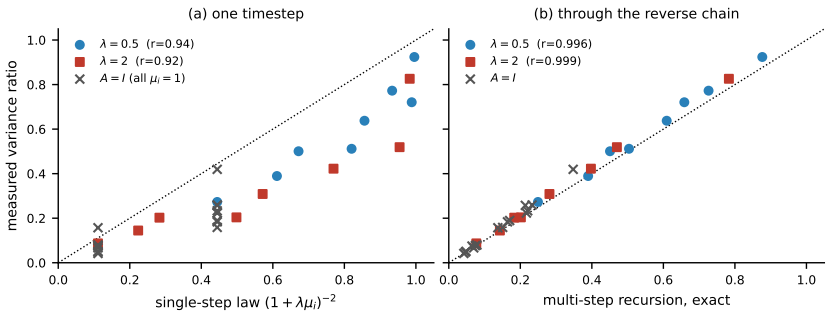



Figure 6: Per-eigendirection variance ratio of a conditional DDPM on a synthetic mixture with a deterministic class target; dotted line is equality. **(a)** Against the single-step law: the ordering is predicted ($r \approx 0.93$) but every point lies below the line, and $A = I$ directions (crosses) that should share one value spread over 0.16–0.42. **(b)** Against the recursion of Proposition 2, no free parameter: $r \geq 0.99$, mean error $\leq 0.033$.

## J Remarks on the scope of the analysis

**Lemma 1** (Error amplification). *Let $\delta_t = \varepsilon - \varepsilon_\theta(x_t, t, y)$. Then exactly $\hat{x}_0 - x_0 = \sqrt{(1-\bar{\alpha}_t)/\bar{\alpha}_t}\,\delta_t$, so $\mathbb{E}\|\hat{x}_0 - x_0\|^2 = \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}\mathbb{E}\|\delta_t\|^2$. The factor diverges as $\bar{\alpha}_t \to 0$; the product need not (at the Bayes optimum for Gaussian $x_0$ it equals $\operatorname{tr}\operatorname{Cov}(x_0 \mid x_t) \to \operatorname{tr}\Sigma$), but for any denoiser with $\mathbb{E}\|\delta_t\|^2 \geq c > 0$ near $\bar{\alpha}_t \to 0$ it grows as $c(1-\bar{\alpha}_t)/\bar{\alpha}_t$. Weighting each sample by $\bar{\alpha}_t$ replaces the factor with $1 - \bar{\alpha}_t \leq 1$.*

*Remark* 4 (Amplification is a property of the estimator). Lemma 1 describes amplification of denoiser error, not unconditional divergence of the population objective. At the Gaussian Bayes optimum, the clean-estimate error remains finite. In our trained models, high-noise residuals are large enough for unweighted coupling to destabilise training; $\bar{\alpha}_t$ weighting controls this amplification (§5.6).

*Remark* 5 (The Loewner strengthening fails). It is tempting to strengthen Theorem 1(iii) to $\operatorname{Cov}(\hat{g}^\star) \preceq \operatorname{Cov}(m)$ on the grounds that $M \preceq I$. That implication is *false* for general positive semi-definite $\Sigma$: with $M = \operatorname{diag}(1, 0.1)$ and $\Sigma = \left(\begin{smallmatrix} 1 & 0.9 \\ 0.9 & 1 \end{smallmatrix}\right)$, the matrix $\Sigma - M\Sigma M$ has eigenvalues $\{-0.454,\ 1.444\}$ and is indefinite. The eigendirectional identity and trace bound remain valid.

*Remark* 6 (Relation to ridge regression). Part (ii) of Theorem 1 is ridge regression and its target-independent covariance is immediate from the affine estimator. What the diffusion setting adds is the rest: no admissible target removes the variance term, an implementation's nominal $\lambda$ is not the theorem's (Appendix L), and the reverse chain reshapes the contraction direction by direction (Proposition 2), which is why the single-step factor is not the sampled one.

## K Latent diffusion

A latent model estimates $\hat{z}_0$ and maps it to signal space through a decoder $\phi$. Applying the loss after decoding gives $\lambda\mathbb{E}\|A\phi(\hat{z}_0) - A\tau_y\|^2$. The following result separates the exact covariance decomposition from the affine contraction law and its local nonlinear approximation.

**Proposition 7** (The variance penalty survives the decoder). *Let $\hat{h} = \hat{h}(z_t, t, y)$ be any measurable clean-latent estimator, $\phi$ any measurable decoder with $\mathbb{E}\|\phi(\hat{h})\|^2 < \infty$, and $\tau_y$ deterministic given $y$. Then at fixed $(y, t)$: **(i)** exactly,*

$$\mathbb{E}\big\|A\phi(\hat{h}) - A\tau_y\big\|^2 = \big\|A\,\mathbb{E}[\phi(\hat{h})] - A\tau_y\big\|^2 + \operatorname{tr}\big(A\ \operatorname{Cov}(\phi(\hat{h}))\,A^\top\big),$$

*so the coupling adds the within-class covariance of the* decoded *estimate to the loss for every decoder, linear or not; the target enters only the first term. **(ii)** If $\phi$ is affine, $\phi(z) = \phi_0 + Jz$, then the coupled latent objective $\mathbb{E}\|\hat{h} - z_0\|^2 + \lambda\mathbb{E}\|A\phi(\hat{h}) - A\tau_y\|^2$ is a ridge problem in $\hat{h}$ with operator $AJ$ and target vector $b = A(\tau_y - \phi_0)$, which need not lie in the range of $AJ$; the derivation of Theorem 1(ii) uses only the quadratic form and goes through unchanged. Its minimiser is*

$\hat{h}^\star = M_J\big(m_z + \lambda J^\top A^\top A(\tau_y - \phi_0)\big)$ *with* $M_J = (I + \lambda J^\top A^\top AJ)^{-1}$ *and* $m_z = \mathbb{E}[z_0 \mid z_t, y]$*, so that for a unit eigenvector* $u_i$ *of* $J^\top A^\top AJ$ *with eigenvalue* $\nu_i$*,*

$$\mathrm{Var}(u_i^\top \hat{h}^\star) = (1 + \lambda\nu_i)^{-2}\ \mathrm{Var}(u_i^\top m_z), \qquad \mathrm{Cov}\left(\phi(\hat{h}^\star)\right) = J\, M_J\, \mathrm{Cov}(m_z) M_J^\top J^\top,$$

*and* $\operatorname{tr}\mathrm{Cov}(\phi(\hat{h}^\star)) \le \|J\|_2^2\ \operatorname{tr}\mathrm{Cov}(m_z)$*. Latent directions in the nullspace of* $AJ$ *are not contracted; every latent direction the read-out* $A$ *can see through the decoder is, by a factor that does not contain* $\tau_y$*.* ***(iii)*** *If* $\phi$ *is* $C^1$ *and the estimator's within-class spread is small relative to the curvature of* $\phi$*, the first-order (delta-method) statement is (ii) with* $J = \nabla\phi$ *at the class-mean latent, and the stationarity condition of the exact nonlinear problem is the fixed point* $\hat{h}^\star = m_z - \lambda\,\nabla\phi(\hat{h}^\star)^\top A^\top A\big(\phi(\hat{h}^\star) - \tau_y\big)$*, a ridge pull along* $\nabla\phi^\top A^\top A$ *toward the target in signal space.*

*Proof.* (i) is the bias–variance decomposition about the deterministic point $A\tau_y$, applied to the random vector $\phi(\hat{h})$; it uses only $\mathbb{E}\|\phi(\hat{h})\|^2 < \infty$. For (ii), substitute $\phi(z) = \phi_0 + Jz$: the coupling becomes $\lambda\mathbb{E}\|AJ\hat{h} - b\|^2$ with $b = A(\tau_y - \phi_0)$. The pointwise first-order condition $(I + \lambda J^\top A^\top AJ)\hat{h} = m_z + \lambda J^\top A^\top b$ gives the minimiser without any preimage of $b$; since only $m_z$ is random, $\mathrm{Cov}(\hat{h}^\star) = M_J\,\mathrm{Cov}(m_z)M_J^\top$, the eigendirectional identity follows as in Theorem 1(iii), and the decoded covariance follows by congruence with $J$. The trace bound uses $\operatorname{tr}(JM_J\Sigma M_J^\top J^\top) \le \|J\|_2^2 \operatorname{tr}(M_J\Sigma M_J^\top) \le \|J\|_2^2 \operatorname{tr}\Sigma$ with $M_J^2 \preceq I$. For (iii), differentiate the pointwise objective $\|\hat{h} - m_z\|^2 + \lambda\|A\phi(\hat{h}) - A\tau_y\|^2$ in $\hat{h}$ and set to zero; the linearisation replaces $\phi$ by its tangent map at the class mean. □

The decoded covariance penalty remains present for every decoder. For an affine decoder, contraction depends on the spectrum of $J^\top A^\top AJ$; directions in the nullspace of $AJ$ remain unconstrained. With a nonlinear decoder, the local approximation can depend on the region occupied by the latent samples. Jointly optimising the decoder provides another possible response to the penalty: reducing its sensitivity along directions of latent spread. The experiments below freeze the decoder.

**Test on a small latent model.** We fitted a VAE to the same beats (convolutional encoder, $K = 32$ latent dimensions, decoder either *linear*, so that part (ii) is exact, or a two-layer MLP, for part (iii); reconstruction RMSE $0.039$ and $0.025$ mV), froze it, and trained a latent DDPM (MLP denoiser, $T = 1000$) on the encoded beats with the point coupling applied to the *decoded* clean estimate, $\lambda\,\bar{\alpha}_t\,\mathrm{mean}\|D\phi(\hat{z}_0) - DT_y\|^2$, the same templates as Table 2; exact identities were enforced on the decoded signal in the loss and at sampling. With the $\varepsilon$-MSE averaged over $K$ dimensions and the coupling averaged over the $1524$ derivative entries, the effective operator is $\tilde{A} = AJ/\sqrt{1524}$ and $\lambda_t = K\lambda(1 - \bar{\alpha}_t)$. Nominal $\lambda$ is therefore not comparable between the two models: the signal-space objective has $\lambda_t = 1536\,\lambda(1 - \bar{\alpha}_t)$ on the operator $D/\sqrt{1524}$, whose largest eigenvalue is $4/1524$, so the coupling strength on the most contracted direction is $\lambda_t\mu_{\max} \approx 4.0\,\lambda(1 - \bar{\alpha}_t)$, against $32\,\lambda(1 - \bar{\alpha}_t) \times 0.0126 \approx 0.40\,\lambda(1 - \bar{\alpha}_t)$ here; latent $\lambda = 10$ matches signal-space $\lambda = 1$, motivating the tested latent weights $\lambda = 1$, 4, and 10. The recursion (4) was run with $\Sigma$ the within-class covariance of the encoded training beats and $M_t = (I + \lambda_t\tilde{A}^\top\tilde{A})^{-1}$, no fitted parameter, and its per-eigendirection variance ratios compared with the measured ratios $\mathrm{Var}(u_i^\top z_{\text{point}})/\,\mathrm{Var}(u_i^\top z_{\text{none}})$ of the sampled latents, $u_i$ the eigenvectors of $\tilde{A}^\top\tilde{A}$ (Figure 7; three seeds per cell).

| decoder, $\lambda$ | arm | feature $\mathrm{MMD}^2$ | raw $\mathrm{MMD}^2$ | diversity | limb (mV) | law: $r$ / MAE |
|---|---|---|---|---|---|---|
| linear, 1 | none | 0.039 ±0.002 | 0.022 ±0.003 | 0.863 ±0.021 | 0.093 | |
| | point | 0.054 ±0.003 | 0.028 ±0.004 | 0.804 ±0.019 | 0.113 | 0.94 / 0.040 |
| | point / exact | 0.052 ±0.002 | 0.028 ±0.004 | 0.804 ±0.019 | 0.000 | |
| linear, 4 | point | 0.091 ±0.003 | 0.041 ±0.004 | 0.693 ±0.024 | 0.087 | 0.96 / 0.105 |
| | point / exact | 0.089 ±0.003 | 0.041 ±0.004 | 0.693 ±0.025 | 0.000 | |
| linear, 10 | point | 0.138 ±0.005 | 0.055 ±0.004 | 0.591 ±0.022 | 0.054 | 0.96 / 0.138 |
| | point / exact | 0.135 ±0.004 | 0.055 ±0.004 | 0.592 ±0.022 | 0.000 | |

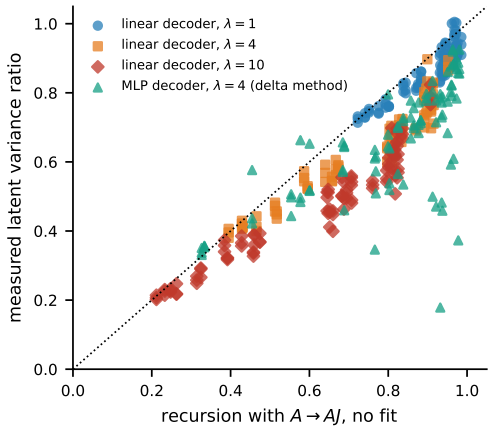



Figure 7: Latent model: measured per-eigendirection variance ratio of sampled latents (point / none) against the recursion of Proposition 2 with $A \to AJ$, no fitted parameter; 32 directions, three seeds per configuration, dotted line equality.

At $\lambda = 10$, the matched effective latent strength, point coupling increases raw $\mathrm{MMD}^2$ by $2.5\times$ and reduces diversity to $0.69$ of the unregularised latent model. The corresponding signal-space values are $3.2\times$ and $0.74$. Across latent weights $1, 4, 10$, raw penalties are $1.3$, $1.85$, and $2.5$, and diversity ratios are $0.93$, $0.80$, and $0.69$.

With the linear decoder, recursion predictions correlate with measured directional ratios at $r = 0.94$–$0.96$, with mean absolute errors $0.040$, $0.105$, and $0.138$. At $\lambda = 10$, predicted relative decoded diversity is $0.785$, versus $0.685 \pm 0.010$ measured. The recursion again underestimates contraction.

The nonlinear decoder is less consistent. At $\lambda = 4$, raw $\mathrm{MMD}^2$ increases $2.0\times$ and diversity falls to $0.75$. Two seeds follow the local recursion ($r = 0.97, 0.94$), but the third has an over-dispersed unregularised baseline ($1.33$ of real diversity, limb residual $1.0$ mV) and weak agreement ($r = 0.35$). A linearisation at the class mean does not describe all sampled states.

Exact identities reduce decoded limb residuals to zero without changing raw distance or diversity to three decimals; feature distance changes by about $3\%$. These small latent models have worse absolute fidelity than the signal-space DDPM (unregularised feature $\mathrm{MMD}^2$ $0.039$ versus $0.003$). They test the decoder-dependent mechanism, not architectural superiority.

## L Coupling schedules, region schedules and inference-time guidance

Let $\rho(t)$ be the nominal coefficient of the simulator loss alongside a noise-prediction MSE averaged over $d$ dimensions. Since

$$\|\hat{\varepsilon} - \varepsilon\|^2 = \frac{\bar{\alpha}_t}{1 - \bar{\alpha}_t} \|\hat{g} - x_0\|^2, \qquad \lambda_t = \frac{d\rho(t)}{\mathrm{SNR}_t},$$

the effective clean-signal weight depends on both the schedule and the loss reductions. For a non-vanishing high-noise denoiser residual, $\rho(t) = O(\mathrm{SNR}_t)$ controls amplification (Lemma 1). Constant $\rho$ yields $\lambda_t \propto 1/\mathrm{SNR}_t$, which grows without bound at high noise. Our $\rho = \lambda\bar{\alpha}_t$ yields $\lambda_t = d\lambda(1 - \bar{\alpha}_t)$, bounded and vanishing at low noise. SNR balancing, $\rho = c\,\mathrm{SNR}_t$, instead gives constant $\lambda_t = dc$. The cutoff applies $\lambda_t = d$ only where $\mathrm{SNR}_t \le \tau$.

Across three seeds, SNR balancing gives diversity $0.58 \pm 0.03$ and about tenfold higher feature $\mathrm{MMD}^2$ than $\bar{\alpha}_t$ weighting. The cutoff preserves similar diversity and nearly halves feature distance (§5.2). At low noise, the posterior has resolved most within-class variance, so a non-vanishing coupling acts directly on that spread. At high noise, the small posterior gain leaves less resolved variance to contract.

**Region-dependent noise schedules.** An exploratory region-dependent noise schedule, adapted from RDDM (Shome et al., 2024) to cardiac phase (a per-index $\bar{\alpha}_t^{s_i}$, so that every region still reaches $\bar{\alpha} \approx 0$ at $t = T$ and sampling stays well posed), dials the diversity of the masked region monotonically in $s$ with the other regions unchanged, but no setting improves $\mathrm{MMD}^2$ beyond seed noise, and on top of the corrected coupling $s = 2$ is destructive ($\mathrm{MMD}^2$ $0.0075 \to 0.1622$) because the coupling's reliability weight is the scalar $\bar{\alpha}_t$ while the masked region carries $\bar{\alpha}_t^s$ (Lemma 1). One caveat on anatomy: the masks were taken from the simulator's model coordinates, in which the R peak sits at $\theta = 0$, while the beat extractor aligns the R peak at index 46 of 128 ($\theta = -0.87$); the window the code labels QRS is therefore the J-point and early-ST segment (Appendix P places its windows around the true R). Changing regional diversity did not improve fidelity in this experiment.

**Inference-time guidance.** The decomposition (2) does not care *when* the pull toward a deterministic target is applied. A plug-and-play guidance step $\hat{x}_0 \leftarrow \hat{x}_0 - \eta A^\top A(\hat{x}_0 - \tau_y)$ is a gradient step on the same quadratic and contracts the estimate by exactly $I - \eta A^\top A$: its eigendirectional variance factor is $(1 - \eta\mu_i)^2$, not $(1 + \eta\mu_i)^{-2}$, so strict contraction requires $0 < \eta\mu_i < 2$ and larger steps expand variance. The recursion (4) applies with $M_t$ replaced by $I - \eta_t A^\top A$; identifying that with the proximal factor $(I + \eta_t A^\top A)^{-1}$ is a first-order approximation. The covariance factor again does not contain $\tau_y$; we have not tested this extension. Classifier-free guidance differs in kind: its reference is the unconditional score, not a fixed target, so no $\mathrm{tr}(A\,\mathrm{Cov}(\hat{g})A^\top)$ term appears, and its documented diversity loss is a sharpening toward the conditional mode, which the decomposition does not describe.

## M Downstream and paired evaluation

SE-Diff's released evaluation compares generated ECGs with paired records using MAE, MSE, RMSE, and Pearson correlation. Our class-conditional generator has no sample-specific ground truth. We instead compare generated beats with the nearest real beat of the same class (precision), real beats with the nearest generated beat (coverage), and generated beats with the real class mean. We subsample each side to 300 beats to match nearest-neighbour sample counts.

We also report TSTR, label fidelity (macro recall of a real-trained classifier on generated beats), and augmentation (AUC after adding 1024 generated beats to real training data). Table 7 uses three generator seeds.

Table 7: Downstream and SE-Diff-style paired evaluation, three generator seeds. Distances in mV RMSE (correlations, all 0.79–0.81, omitted). The reference row is the real test set paired against itself, the TRTR bound, the real-classifier recall on real test beats, and real-only training.

**A. Classification**

| Training set | TSTR | Label fidelity | Augment 1024 real | Augment 3400 real |
|---|---|---|---|---|
| Real reference | 0.846 | 0.636 | 0.829 | 0.880 |
| None | 0.791 ±0.019 | 0.551 ±0.024 | 0.811 ±0.011 | 0.866 ±0.005 |
| Point / soft | 0.732 ±0.021 | 0.552 ±0.019 | 0.819 ±0.009 | 0.858 ±0.004 |
| Spread / exact | 0.729 ±0.015 | 0.517 ±0.012 | 0.807 ±0.007 | 0.860 ±0.007 |

**B. Paired distances (mV RMSE)**

| Training set | Precision | Class mean | Coverage |
|---|---|---|---|
| Real reference | 0.107 | 0.135 | 0.117 |
| None | 0.104 | 0.130 | 0.115 |
| Point / soft | 0.089 | 0.097 | 0.115 |
| Spread / exact | 0.101 | 0.125 | 0.117 |

Point coupling has the smallest distance to the class mean (0.097 mV, versus 0.135 for real beats) and the best nearest-real precision distance. These measures can favour concentration. Raw and

feature $\mathrm{MMD}^2$ and TSTR detect the resulting loss of variation, whereas nearest-generated coverage changes little at this sample size. Label fidelity is similar for point coupling and the baseline ($0.55$), and lower for spread/exact ($0.52$). Adding generated beats does not improve the real-data classifier for any arm in these runs.

## N MIMIC-IV-ECG REPLICATION AND TRANSFER

MIMIC-IV-ECG v1.0 supports two tests: retraining all three principal arms with newly fitted templates, and evaluating PTB-XL-trained generators under a change of dataset. The first tests replication; the second tests transfer.

**Training on MIMIC.** We derive superclass labels from Marquette 12SL statements in `machine_measurements.csv` using keyword rules. We retain records assigned to exactly one superclass and ignore hedged statements such as "cannot rule out". From $447{,}207$ single-superclass records, we select $600$ records per class with one record per subject. Resampling from $500$ to $100$ Hz and applying the beat pipeline yields $2935$ beats ($579$–$596$ per class). The $85/15$ split is subject-disjoint, with $2494$ training and $441$ test beats.

We refit Gabor templates and the 32-member bank on MIMIC training beats (template RMSE $0.009\,\mathrm{mV}$), then train each principal arm with three seeds. The CNN real-data reference uses $1024$ MIMIC training beats. Its AUC is $0.74$, compared with $0.83$ on PTB-XL. Keyword-derived machine labels and the small sample limit comparisons of absolute scores across datasets; the controlled result is the within-MIMIC comparison between arms.

| coupling | feature $\mathrm{MMD}^2$ | Raw $\mathrm{MMD}^2$ (ratio) | diversity | limb (mV) | TSTR AUC |
|---|---|---|---|---|---|
| real (1024 beats) | | | | | 0.740 ±0.012 |
| none | 0.0034 ±0.0015 | 0.0069 ±0.0027 (1.0) | 0.866 ±0.013 | 0.083 | 0.698 ±0.015 |
| point / soft | 0.168 ±0.007 | 0.0222 ±0.0037 (3.2) | 0.628 ±0.004 | 0.071 | 0.672 ±0.002 |
| spread / exact | 0.0065 ±0.0012 | 0.0074 ±0.0029 (1.1) | 0.873 ±0.011 | 0.000 | 0.698 ±0.001 |

On MIMIC training data, point coupling increases raw $\mathrm{MMD}^2$ by $3.2\times$, reduces pooled diversity to $0.72$ of baseline, and lowers TSTR by $0.026$. Spread/exact matches baseline raw distance, diversity, and TSTR ($0.698$ for both), but has $1.9\times$ its feature distance. MIMIC's single real-versus-real reference is $-0.0016$ in feature space and $0.0002$ in raw space. We do not divide by these references.

**Cross-database evaluation.** We also evaluated the PTB-XL-trained generators of Table 2 against the same $2{,}935$ MIMIC beats. The reference row is PTB-XL's own held-out real beats against MIMIC , which measures the real-data domain gap.

Table 8: PTB-XL-trained generators against MIMIC-IV-ECG real beats. $\mathrm{MMD}^2$ under the feature map and on the raw signal, with the ratio to the PTB-XL→MIMIC domain gap in brackets; pooled within-class diversity relative to MIMIC. Three generator seeds; the MIMIC real-vs-real split value is $0.0002$ (feature) and $0.0001$ (raw) at this sample size, so ratios to it are not informative.

| | feature $\mathrm{MMD}^2$ | raw $\mathrm{MMD}^2$ | Diversity / MIMIC | TSTR on MIMIC |
|---|---|---|---|---|
| PTB-XL real 1024 beats | 0.0309 [1.0] | 0.0116 [1.0] | 1.10 | 0.692 ±0.005 |
| none | 0.0307 ±0.0016 [1.0] | 0.0169 ±0.0016 [1.5] | 0.96 ±0.02 | 0.654 ±0.006 |
| point / soft | 0.1405 ±0.0014 [4.5] | 0.0257 ±0.0008 [2.2] | 0.72 ±0.02 | 0.662 ±0.002 |
| spread / exact | 0.0338 ±0.0048 [1.1] | 0.0186 ±0.0005 [1.6] | 0.91 ±0.02 | 0.624 ±0.009 |

Against MIMIC data, PTB-XL-trained models retain their distributional ordering. Feature distances relative to the real PTB-XL–MIMIC gap are $1.0$ for the baseline, $1.1$ for spread/exact, and $4.5$ for point coupling. All generated raw distances exceed that gap ($1.5$–$2.2\times$), and diversity ratios to MIMIC are $0.96$, $0.91$, and $0.72$, respectively.

**Cross-database classification.** TSTR on MIMIC is $0.654 \pm 0.006$ for the baseline, $0.662 \pm 0.002$ for point coupling, and $0.624 \pm 0.009$ for spread/exact. Training on $1024$ real PTB-XL beats gives $0.692 \pm 0.005$; chance performance is $0.502$. Point coupling has higher AUC on MI ($0.699$ versus $0.660$), CD ($0.702$ versus $0.649$), and HYP ($0.574$ versus $0.551$), but lower AUC on NORM ($0.717$ versus $0.759$) and STTC. Its macro advantage, $0.008 \pm 0.006$, remains uncertain with three generator seeds. Suppressing dataset-specific variation could aid transfer, but this experiment does not identify that mechanism.

## O MULTI-BEAT WINDOWS

Phase-aligned single beats remove RR variability and beat-to-beat relations. We repeat the principal comparison on $4$ s, 12-lead windows ($400$ samples at $100$ Hz), starting $0.25$ s before an R peak. Up to two non-overlapping windows per record yield $7848$ windows from the same $4000$ records, split patient-disjoint into $6671$ training and $1177$ test windows. Each extracted window contains an average of $4.6$ beats. We train the same UNet for $100$ epochs with a comparable gradient-step budget.

For point coupling, we extract a $0.7$ s segment of $\hat{x}_0$ around each stored R position of the real training window, resample it to $128$ phase points, and match its derivative to the class template. The bank, nominal weight, and $\bar{\alpha}_t$ weighting match the beat experiments. Neither window arm includes an inter-lead loss. Anchoring the loss to real R positions constrains morphology and may also affect generated timing.

Table 9 reports, over three generator seeds, the raw-signal $\mathrm{MMD}^2$ against real test windows (flattened $12 \times 400$; real-vs-real split value $0.0002$), pooled within-class diversity, the RR-interval distribution recovered by running the same R detector on generated lead II (Kolmogorov–Smirnov distance to the real RR distribution, mean and sd, beats per window), and a train-synthetic/test-real classifier trained on $1024$ generated windows.

Table 9: Multi-beat $4$ s windows, patient-disjoint, mean$\pm$sd over three generator seeds. Real windows: RR $0.796 \pm 0.193$ s, $5.15$ beats per window, real-vs-real $\mathrm{MMD}^2$ split value $0.0002$; a classifier trained on $1024$ real training windows scores $0.806$ AUC. The point coupling reproduces the beat-level penalty on distributional distance and diversity, the measured RR summaries differ little between arms, and the classifier is tied.

| Measure | None | Point |
|---|---|---|
| Raw $\mathrm{MMD}^2$ | $0.0195 \pm 0.0009$ | $0.0546 \pm 0.0033$ |
| Raw ratio | 1.0 | 2.8 |
| Diversity | $0.703 \pm 0.005$ | $0.567 \pm 0.004$ |
| RR KS | $0.148 \pm 0.007$ | $0.148 \pm 0.009$ |
| RR mean $\pm$ sd (s) | $0.867 \pm 0.347$ | $0.866 \pm 0.340$ |
| TSTR AUC | $0.765 \pm 0.007$ | $0.774 \pm 0.020$ |

The beat-level distributional penalty persists in windows: raw distance increases $2.8\times$ and pooled diversity falls from $0.70$ to $0.57$. Both arms have RR KS distance $0.148$ and similar over-dispersion (approximately $0.35$ s, versus $0.19$ s in real windows), with $4.5$–$4.6$ detected beats per window versus $5.15$ in real data. TSTR differs by $0.009 \pm 0.021$ in favour of point coupling.

These measurements do not prove that timing is unaffected. The extraction operator depends on each training window's R positions, whereas Theorem 1 assumes a fixed operator. Anchored waveform matching may penalise timing changes, and the reported RR summaries may miss small effects. The experiment tests within-beat template matching in rhythm-preserving windows; it does not test a loss that constrains RR intervals or beat-to-beat relations directly.

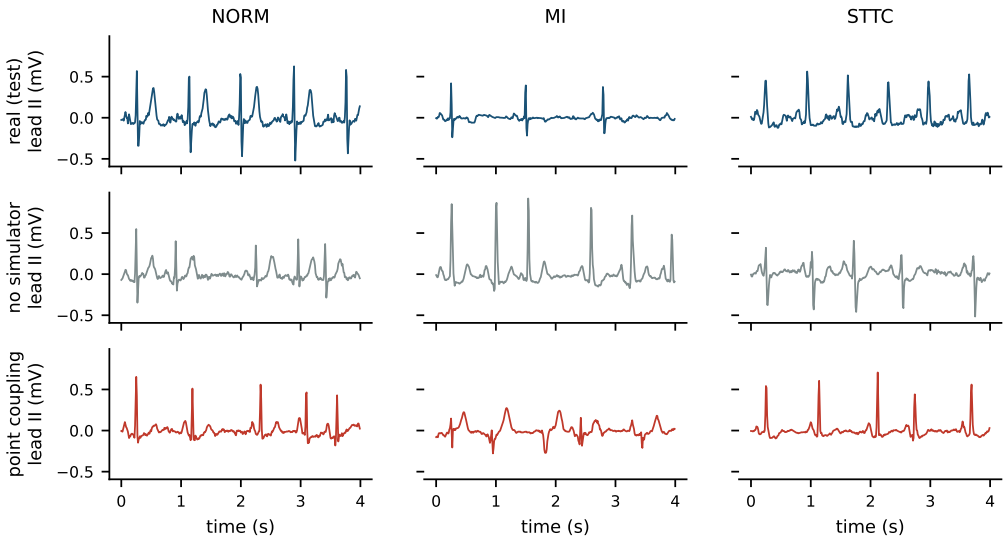



Figure 8: Lead II of 4 s windows (patient-disjoint test set, and seed-0 samples of each arm): real (top), unregularised model (middle), point coupling (bottom). The examples illustrate morphology and beat timing; Table 9 reports aggregate waveform and RR statistics.

## P CLINICAL MEASUREMENTS

We apply identical fiducial detectors to real beats and samples from the three principal arms, using three generator seeds. QRS onset and offset use spatial velocity around the R peak; P onset and T end use lead II at 20% and 10% of wave amplitude. Intervals are fractions of the cardiac cycle because phase normalization removes absolute timing. QTc cannot be defined without RR intervals. Learned-space measures use the real-trained beat classifier's 128-dimensional penultimate embedding: Fréchet distance, $\mathrm{MMD}^2$, and $k$-NN precision and recall (Kynkäänniemi et al., 2019).

Table 10: Clinically interpretable measurements, mean over three generator seeds. Intervals in fractions of the cardiac cycle, median [IQR], with the Kolmogorov–Smirnov distance to the real distribution; R amplitude in mV; frontal QRS axis in degrees for NORM; Einthoven residual RMS in mV; the last two rows are in the learned embedding, where the real-versus-real reference is FD 1.14, recall 0.95, precision 0.94.

| | real | none | point / soft | spread / exact |
|---|---|---|---|---|
| PR | 0.195 [0.023] | 0.198 [0.023] | 0.195 [0.016] | 0.195 [0.023] |
| KS vs real | — | 0.03 | 0.12 | 0.03 |
| QRS | 0.102 [0.016] | 0.102 [0.016] | **0.109** [0.023] | 0.102 [0.023] |
| KS vs real | — | 0.03 | **0.31** | 0.05 |
| QT | 0.492 [0.086] | 0.497 [0.090] | 0.503 [0.097] | 0.497 [0.094] |
| KS vs real | — | 0.05 | 0.08 | 0.05 |
| R amplitude V4 / V5 | 1.10 / 1.17 | 1.07 / 1.11 | **0.87** / **0.94** | 1.06 / 1.09 |
| R monotone V1→V4 | 0.72 | 0.74 | 0.78 | 0.74 |
| QRS axis, NORM | 27.6 [44.3] | 29.0 [50.5] | 32.3 [36.1] | 30.4 [53.4] |
| Einthoven residual | 0.0001 | 0.0030 | 0.0026 | 0.0000 |
| Fréchet distance | 1.14 | 0.93 | **2.79** | 1.32 |
| recall / precision | 0.95 / 0.94 | 0.90 / 0.96 | **0.75** / 0.98 | 0.88 / 0.97 |

**Morphology.** Under point coupling, median QRS duration increases from 0.102 to 0.109 of the cycle, with KS distance 0.31; the other generated arms are within 0.05. Precordial R amplitudes are lower (V4: 1.10 to 0.87 mV), and frontal-axis IQR narrows from approximately 44° to 36°. PR spread also decreases while its median remains similar. Smooth derivative-matched templates

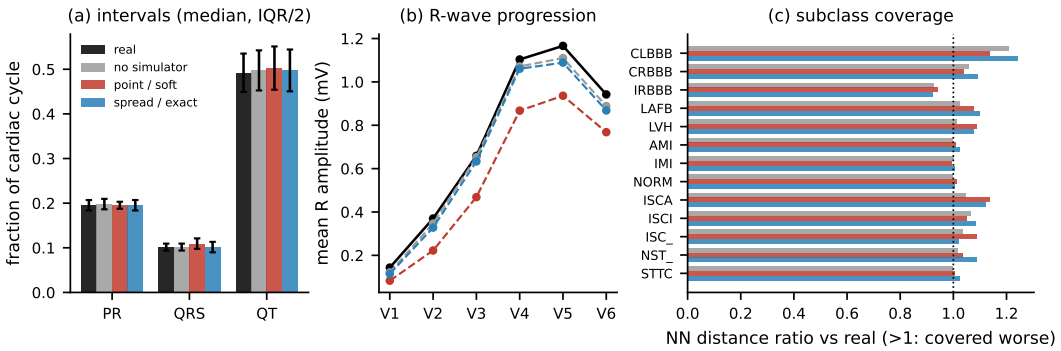



Figure 9: **(a)** PR, QRS and QT as fractions of the cycle (median, half-IQR bars). **(b)** R-wave progression: mean R amplitude V1–V6. **(c)** Coverage of PTB-XL diagnostic subclasses: nearest-neighbour distance from real test beats of a subclass to generated beats of its superclass, relative to the same distance to real training beats of matched size.

and averaging across beat-to-beat variability are possible explanations. Spread/exact brings these measures closer to the unregularised model and eliminates the Einthoven residual.

**Subclass coverage.** For the 13 subclasses with at least eight test beats, coverage ratios range from 0.92 to 1.24 without a consistent arm ordering (Figure 9c). Nearest-neighbour distance in 1536 dimensions is insufficiently informative here to establish coverage of rare morphologies.

**Learned-space recall.** Precision is 0.96–0.98 for all generated arms. Recall is 0.95 for real-versus-real comparisons, 0.90 without coupling, 0.75 for point coupling, and 0.88 for spread/exact. Fréchet distances are 0.93, 2.79, and 1.32, versus a real-versus-real reference of 1.14. These estimates support reduced coverage under point coupling within this embedding and neighbourhood estimator; they do not establish clinical validity of individual samples.

## Q GENERATED BEATS

Figure 10 overlays twenty lead-II beats per class from real data and each principal arm. For NORM, within-class standard deviation is 0.062 mV in real beats, 0.057 without coupling, 0.047 with point coupling, and 0.056 with spread/exact. Point coupling gives the narrowest spread in four of five classes. In CD, spread/exact is narrower by 0.002 mV (0.061 versus 0.063), within the uncertainty of a 79-beat estimate.

The 12-lead NORM means in Figure 11 differ from the real mean by 0.0130 mV without coupling, 0.0117 for point coupling, and 0.0121 for spread/exact. These similar means are compatible with different spreads. Feature-space displacement can reflect changes in per-lead standard deviation rather than a displaced waveform mean. The moment-matching arm, whose mean term targets the template bank, is not plotted.

## R DENOISER COVARIANCE ON FIXED INPUTS

Theorem 1 concerns the denoiser's estimate $\hat{g}(x_t)$ when $x_t$ is drawn from the forward process at a fixed timestep; the diversity numbers of the main text concern terminal samples of the reverse chain, whose input distribution moves with the coupling. To separate the two we feed the *same* forward-noised real test beats (shared noise) to the unregularised and the point-coupled networks (seed 0, $\lambda = 1$) and compare the within-class variance of $D\hat{g}$ along the eigendirections of $D^\top D$ at fixed $t$ with the corresponding ratio of reverse-chain samples. The single-step law at that $t$ uses the effective weight $\lambda_t$ of §3.2.

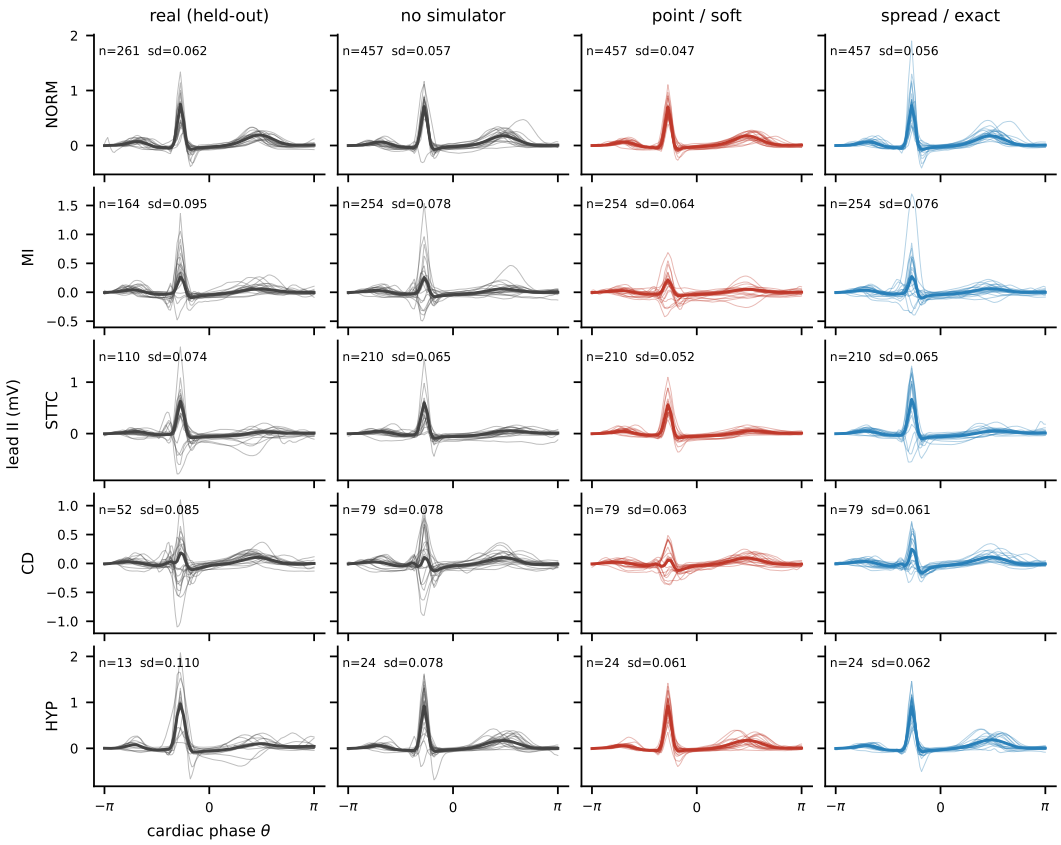



Figure 10: Lead-II beats by diagnostic superclass (rows) and arm (columns): twenty beats thin, class mean bold, within-class sd printed. Point coupling has the narrowest spread in four of five classes (the CD difference is $0.002$ mV); the corrected coupling restores most of the real spread.

| $t$ | $\bar{\alpha}_t$ | $\lambda_t$ | Pooled sd ratio | Mean eigen-ratio | Law $(1+\lambda_t\mu_i)^{-2}$ | $r$ |
|---|---|---|---|---|---|---|
| 20 | 0.994 | 0.006 | 0.999 | 0.959 | 0.976 | 0.97 |
| 100 | 0.895 | 0.106 | 0.986 | 0.683 | 0.716 | 0.99 |
| 300 | 0.394 | 0.611 | 0.939 | 0.385 | 0.351 | 0.97 |
| 600 | 0.026 | 0.982 | 0.914 | 0.345 | 0.275 | 0.98 |
| 900 | 0.000 | 1.008 | 0.824 | 0.427 | 0.271 | 0.99 |
| reverse-chain samples | | | 0.741 | | | |

At fixed forward-process inputs, directional variance ratios correlate with the single-step law at $r \geq 0.97$ at every tested timestep. The mean eigen-ratio weights directions equally; the pooled ratio weights them by variance. At high noise, the denoiser contracts less than the population law, yet reverse-chain samples reach pooled standard-deviation ratio $0.741$, below every fixed-input ratio. This is consistent with additional contraction during sampling, but does not isolate its cause. The reverse chain visits states not covered by this probe, so denoiser errors there remain a possible explanation. The experiment uses one seed and one coupling weight.

## S Reproducibility

**Evaluation details.** $\mathrm{MMD}^2$ uses an RBF kernel with the median-heuristic bandwidth computed on the pooled standardised sample, the unbiased U-statistic, and a random subsample of $400$ beats from each side; both sides are standardised feature-wise by the *real* side's mean and standard deviation. The feature map is, per lead, the mean, standard deviation and peak-to-peak over phase plus six

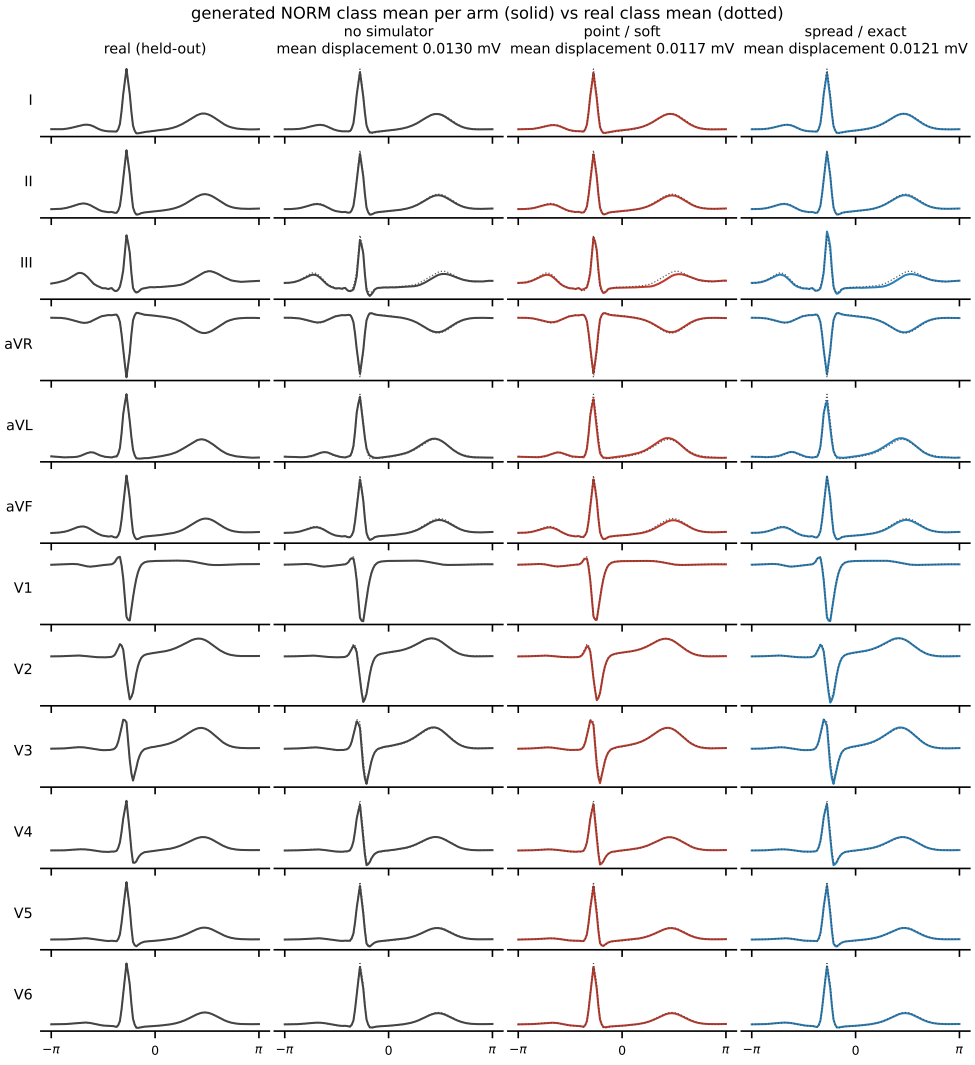



Figure 11: Generated NORM class mean per arm (solid) against the real NORM class mean (dotted), with the RMSE displacement over 12 leads. The three arms displace the mean by 0.012–0.013 mV, alike and a fifth of the within-class sd; what differs between arms is spread.

log-power bands of the rFFT (108 statistics); the raw map is the flattened $12 \times 128$ beat. The real-vs-real reference splits the 600 held-out beats into halves; it is a single draw from the null and is reported for orientation only. The null calibration of Appendix A (`mmd_null.py`) instead fixes one bandwidth per feature map from the real beats alone, uses matched 300-vs-300 counts for every comparison, draws 500 random real-vs-real splits, and scores each arm with 20 random draws per seed. Pooled diversity is $\sqrt{\sum_c \sum_{i \in c} \|x_i - \bar{x}_c\|^2 / \sum_c (n_c - 1)\, d}$ over classes with $n_c \geq 2$, for generated and real beats separately, reported as a ratio. Templates are fitted to the mean beat of a separate cohort of 200 records per class: the first usable records of each superclass under a fixed permutation of all of PTB-XL (`templates.py`, seed 1; one record-mean beat each, baseline removed on the first 7 samples), the bank holding the per-record fits of the first 32. This cohort is drawn independently of the generator's 4000 records and their splits, so overlap is a matter of chance; `template_overlap.py` counts it. Of the 1000 cohort records, 163 lie in the patient-disjoint training set, 28 in its test set and 38 belong to test patients; of the 160 bank members, 6 are test records and 8 belong to test patients (1–2 per class). A test record enters a class template with weight $1/200$; its influence through the bank is not bounded by $1/32$, because a nearest-member target is used as often as it is the closest member and the moment targets depend on the bank statistics. `templates.py --exclude-records` therefore fits a cohort with every test patient removed

(738 records excluded; `runs/tmpl_gabor_bank_clean.npz`, template RMSE $0.0105$ against $0.0104$ mV); replaying its selection confirms zero overlap with the patient-disjoint test set, by record and by patient, for the 1000 cohort records and the 160 bank members (`template_overlap.py --exclude-records`). Every template-using patient-disjoint arm is trained on that bank under the same protocol and seeds (`run_clean.sh`, `runs/psplit_clean`); these clean-cohort arms are the ones reported in Tables 2 and 3 and Figure 1b. The same arms trained on the original cohort (`runs/psplit_seeds`) are the sensitivity comparison of Table 4 in Appendix B; the record-split experiments and the remaining controls use the original cohort. Class labels are PTB-XL diagnostic superclasses of records with a single superclass at likelihood $\geq 50$. No hyperparameter was selected on the test set; all arms use the fixed settings above and one nominal $\lambda$ except where a sweep is reported.

The generative model is a 1-D UNet with class conditioning, 1000 diffusion steps, linear $\beta$ schedule, AdamW at $2 \times 10^{-4}$, cosine learning-rate decay, EMA $0.999$, batch 64, trained on a single consumer GPU. Every arm shares architecture, data split, schedule and budget; only the coupling differs. The released-code checks in §5.6 call the repository's loss functions without modification. The TSTR classifier of §5.7 is a three-block 1-D CNN (widths 32–128, kernel 7, stride 2, global average pool, dropout $0.3$) trained 60 epochs with class-balanced cross-entropy, AdamW at $2 \times 10^{-3}$; the real-data upper bound uses $1024$ real training beats drawn at random so that the comparison is not confounded by sample size. The 1-D ResNet of Appendix B (`tstr_seeds.py --clf resnet`) replaces the body by a stride-2 stem and three stages of two residual blocks (kernel 5, widths 32/64/128, stride 2 at the second and third stage) under the same head, optimiser, epochs and seeds.

**Multi-beat windows** (Appendix O) use the same UNet on $12 \times 400$ inputs, batch 32, 100 epochs, same optimiser, schedule and EMA; windows start 25 samples before a detected R peak (Pan–Tompkins-style detector on lead II), the beat-wise coupling slices $[p - 25, p + 45)$ around each stored R position $p$ of the real window and resamples it linearly to 128 points; the RR statistics run the same detector on generated lead II in mV; the window classifier is the TSTR classifier above, which is length-agnostic through its global pooling.

**MIMIC-IV-ECG training** (Appendix N): `sediff.py` with `--data mimic_ecg.npy --n-records 2990` and `--templates tmpl_gabor_bank_mimic.npz` (fitted by `templates.py` on the MIMIC training beats), otherwise the settings of Table 2; the TSTR bound draws its real beats from the MIMIC training split (`tstr_seeds.py --real-data`).

**Latent model** (Appendix K, `sediff_latent.py`): VAE with a three-layer strided convolutional encoder, $K = 32$, KL weight $10^{-3}$ per signal entry, 80 epochs, Adam $10^{-3}$; decoder linear ($1536 \times 32$ affine map) or MLP ($32 \to 256 \to 512 \to 1536$, SiLU); latents standardised by their global sd, which is folded into $J$. Latent DDPM: MLP denoiser (width 256, sinusoidal time and class embeddings), $T = 1000$, linear $\beta$, 400 epochs, batch 64, AdamW $10^{-3}$, EMA $0.999$; 1024 samples per arm, decoded to mV; the record split of Table 2. The Jacobian for the MLP decoder is the mean of the exact Jacobians at the five class-mean latents.

**Factorial cells** (§5.5): `sediff.py` with `--sim-mode point --inter off` (morphology term only), `--lam-sim 0 --inter soft` (inter-lead term only) and `--sim-mode point --inter exact` (the exact-identity control), six generator seeds each, otherwise the settings of Table 2; `run_factorial.sh` launches the cells on the original cohort and `run_clean.sh` on the clean cohort, and `factorial_summary.py --dir runs/psplit_clean` computes the lead-group diversities, normalized residuals and matched per-seed contrasts of Table 3 from the saved samples. Every run file records its complete argument list. Under the per-run protocol (bandwidth estimated per comparison), the clean-cohort cells read $0.0060$, $0.0142$, $0.0116$ and $0.0189$ raw $\mathrm{MMD}^2$ (none, morphology only, inter-lead only, both) and the raw interaction is $-0.0009$ $[-0.0019, +0.0000]$, the same sign as under the shared kernel (original cohort: $0.0060$, $0.0141$, $0.0115$, $0.0188$ and $-0.0009$ $[-0.0016, -0.0002]$). TSTR values for the patient-disjoint arms in Tables 2 and 3 and in §5.7 and Appendix B come from one run of `tstr_seeds.py` over `runs/psplit_clean` (`runs/tstr_psplit_clean.json`), five classifier seeds per generator seed, and the none/exact control from its own run over the unregularised samples with exact identities applied. The original-cohort TSTR values of Table 4 come from the runs over `runs/psplit_seeds` and the point / exact control listed in `clean_compare.py`. The template-free baseline is scored in both: $0.774 \pm 0.023$ in the clean-cohort run and $0.770 \pm 0.017$ in the original-cohort run, a difference within one classifier-seed sd; an earlier run of the three

principal arms differs by at most $0.005$ in the means. The real-data reference and the per-class real row of Appendix B come from one set of five classifiers trained by the same routine as the arms (`tstr_seeds.py --real-only`); a repeated real-data fit differed by about $0.01$ in macro AUC, an observation about the reference fit only. The uncertainty of each arm's TSTR is the measured sd over its generator seeds in Tables 2 and 3. `make_tables.py` regenerates the rows of Tables 2 and 3 from the result files with one rounding rule, and `clean_compare.py` those of Table 4. The code will be publicly available upon acceptance.

**Seeds and coefficients.** Generator seeds are $0$–$5$ for the patient-disjoint arms of Table 2 and the factorial cells of §5.5 (three for moment matching) and $0, 1, 2$ elsewhere (seed $0$ is the seed of the single-seed sweeps); classifier seeds are $0$–$4$ for TSTR and $0$–$2$ elsewhere. The effective per-step coupling weight implied by the loss reductions is $\lambda_t = (1536/1524)\,\lambda_{\mathrm{sim}}(1-\bar{\alpha}_t)$ on every lead, plus $(1536/762)\,\lambda_{\mathrm{inter}}(1-\bar{\alpha}_t)$ on the six limb leads under the soft inter-lead residual; the $\lambda$-sweep of §5.1 set $\lambda_{\mathrm{inter}} = \lambda_{\mathrm{sim}}$. The SE-Diff functions audited in §5.6 are those of commit `dd0a476d`.

## AI USE STATEMENT

We used ChatGPT for language editing and LaTeX table formatting, and Claude (Anthropic) as a coding and analysis assistant: it drafted and ran the evaluation scripts for the factorial experiment of §5.5, the shared-kernel evaluation with matched draws, the per-seed contrasts and intervals, and the reproducibility package.

## REPRODUCIBILITY STATEMENT

Section 3 and Appendices J and K state the theoretical assumptions and proofs. Appendix S specifies preprocessing, loss reductions, architectures, evaluation procedures, and seeds. Dataset splits and additional experimental settings appear with their corresponding appendix results.