STRUCTURE LIVES IN THE REPRESENTATION: PHASE-FIELD DIFFUSION FOR ROBUST ECG DELINEATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Under noise, the masks of neural ECG delineators fragment and misorder, breaking the $P \to QRS \to T$ sequence that interval measurement presupposes. Point-level F1, the standard metric, overstates the usable fiducials. The standard remedy, rule-based post-processing, reassembles detected waves but cannot synthesize missed ones. We replace repair with generation and move structure into the learning target. We train a conditional diffusion prior over phase fields: each wave class occupies a channel whose value tracks within-wave phase. The target is continuous and invertible, encoding presence, boundaries, and order, not enforcing them afterwards. The conventional one-hot, a flat label with no within-wave position, is structurally fragile. Under the harshest external-test noise, even trained with the field's own loss, it reaches rules only by selecting among samples. Its single draw reads 0.794 against rules' 0.863. Averaging them collapses at the decision-boundary readout of established conventions. The phase field needs no such crutch. A single draw lands in rules' seed band, clearing rules at the reported pairing (0.888); its gain over the matched one-hot is positive at every seed. The same selection lifts harshest-condition legality to 0.994, above the one-hot's 0.980, and averaging stays legal without decode tuning. Legality is not bought by deleting waves: the field misses fewer ground-truth P waves than rules, attains the best F1, and trails rules on intervals by milliseconds. Under the harshest noise, generation replaces rule repair; structure steers generation.

1 INTRODUCTION

Electrocardiogram (ECG) delineation translates a continuous signal into structured output: per-sample labels of the boundaries of P waves, QRS complexes, and T waves. Downstream interval measurement and diagnostic reasoning consume this output (Kligfield et al., 2007; Stracina et al., 2022). The pipeline presupposes that every beat follows the fixed sequence $P \to QRS \to T$. Noise breaks the presupposition: electrode motion, baseline wander, and muscle artifact make the masks of neural delineators *fragment* (a wave splits or vanishes) and *misorder* (missing, spurious, or out-of-order runs), and the sequence becomes uninterpretable (Fig. 1a).

Illegality is not cosmetic: intervals spanning fragmented or misordered waves are silently dropped or mis-measured, and point-level F1, the standard metric, is nearly blind to this failure. We make structure itself the object of measurement: a signal is *legal* if the run sequence of its mask parses under the grammar $(P\,?\;QRS\;T)^{+}$, and *legality* is reported as the *worst cell* over a 12-condition noise grid. On our external test set, a capacity-matched discriminative U-Net anchor under the harshest noise (electrode motion at 0 dB, em@0) still reaches 0.827 point-level F1 while its legality is only 0.544.

The standard remedy for illegal output is rule-based post-processing: merge, filter, and reassemble detected runs anchored on QRS (Pan & Tompkins, 1985; Di Marco & Chiari, 2011). Rules can reshape detected waves but cannot synthesize missed ones; repair is capped by detector recall (quantified in Sec. 4.3). We advocate replacing repair itself: **generation instead of rules**. Train a conditional diffusion prior on the same anchor backbone to model the distribution of legal masks directly; sampling is synthesis, and missed waves can be generated without being detected (Fig. 1b). Under the same harshest condition, rules lift legality from 0.544 to 0.863, and a diffusion prior with a one-hot target reaches 0.980.

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.

This one-hot ceiling, however, rests on sampling: trained with the same x_0-MSE loss as the field, a single draw still sits below rules (0.794 vs. 0.863), and averaging collapses legality under every tested loss and sampler family at the decision-boundary readout both conventions share (Appendix P). The diagnosis is structural: **the one-hot step target carries no within-wave position; structure is not in the representation.** The prescription is a continuous **phase field**: each wave class is carried by a dedicated channel whose value rises and falls with within-wave phase, a C^0-continuous, invertible target that encodes wave *presence*, *boundaries*, and *order* into the learning target itself. Under the field target, a single draw sits in rules' seed band at the harshest cell (0.888 vs. 0.863 at the reported pairing), $+3.2$ to $+11.4$ pp over the matched one-hot at every seed; the same selection lifts it to 0.994, above the one-hot's 0.980 (Fig. 1c): **structure lives in the representation, not in sample selection.**

Legality is not bought by deleting waves either: the field misses fewer ground-truth P and T waves than rules, attains the best F1 of the four variants, and trails rules' interval median errors by milliseconds (Secs. 4.3, 4.5). Our contributions:

1. **A structural legality metric and a three-layer evaluation protocol** (Sec. 3.1, 4.1): "is the mask usable under noise" is separated from point-level F1 and measured directly (structure, usability, precision).

2. **Generation replaces rule post-processing** (Sec. 3.3, 4.2–4.3): a diffusion prior trained with in-model noise corruption restores legality (+45.0 percentage points [pp] over the bare anchor and +13.1 over rule post-processing under the harshest condition, CIs excluding zero), and the wave-deletion audit quantifies the rules' repair cap.

3. **A continuous phase-field target** (Secs. 3.2, 4.4): we diagnose the structural defect of one-hot steps and propose a C^0-continuous, invertible field target, with the failure axes separated. Generation supplies the legality ceiling (0.994 vs. 0.682 for deterministic regression; Appendix J); the one-hot single draw sits below rules under every tested loss (Appendix N); and the representation alone supplies the selection-free operators, where the field succeeds at every seed and under averaging while the one-hot fails both under every tested loss and sampler family at the decision-boundary readout (Appendix P).

4. **Full-dimension evaluation and open source** (Sec. 4, appendices): an external QTDB test with revised labels on a 12-condition grid, paired bootstrap CIs, K/τ/tolerance/seed/channel-held-out scans, and real atrial fibrillation (AF); code, configurations, and evaluation scripts will be released.

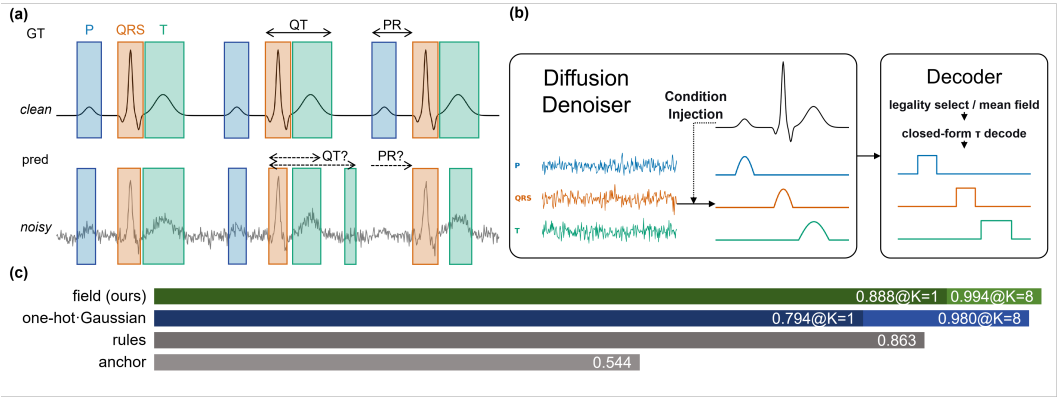


Figure 1: Phenomenon, pipeline, and results. Colors: P blue, QRS orange, T green. (a) Under noise the prediction ($pred$) fragments and misses waves, leaving QT/PR unmeasurable (?) against the ground truth (GT; $clean$: noise-free reference). (b) A diffusion denoiser U_θ conditioned on the corrupted input turns per-channel Gaussian noise into K per-class phase fields (one-hot contrast: Fig. 4); the closed-form τ decode yields per-candidate masks for legality selection or a single mask after averaging. (c) Legality at em@0, every variant's worst cell (Sec. 4.2), for both targets at $K{=}1/8$ (same x_0-MSE loss): a single field draw clears the rules bar; the one-hot's sits below it.

2 RELATED WORK

ECG delineation. The classical pipeline cascades a QRS detector with rule- or wavelet-based analysis (Pan & Tompkins, 1985; Martínez et al., 2004); deep learning recast the task as per-sample segmentation, U-Net family models mainstream (Moskalenko et al., 2020; Ronneberger et al., 2015). Noise robustness is usually addressed by augmentation training (Jimenez-Perez et al., 2024; 2021), our external baseline's configuration; our critique targets structural preservation: a capacity-matched augmented model stays structurally fragile (worst-cell legality 0.488–0.544).

Three destinations for structure: post-processing, structured inference, output representation. The first is *after-the-fact repair*: run merging and grammar filtering in ECG (Di Marco & Chiari, 2011), reorganization only, never creation. The second is *structured inference*: conditional random fields (Zheng et al., 2015), energy-based structured prediction (Belanger & McCallum, 2016), automaton-constrained decoding (Deutsch et al., 2019) (our legality score's finite-candidate form); it buys consistency but only reorganizes what the scores support; we learn none. The third is *encoding structure into the output representation*: distance-field regression in instance segmentation (Bai & Urtasun, 2017; Schmidt et al., 2018; Ma et al., 2020) and phase/progress regression in action segmentation (Dwibedi et al., 2019; Wei & Ye, 2022) show continuously parameterized targets beat per-point labels; generative distance fields are established in image segmentation (Bogensperger et al., 2023; 2025), and the route's *temporal* generative version poses its own question: *whether samples can be averaged tracks the output representation* (Appendix C; at the decision-boundary readout and within the measured regime, Appendices P and Q). The one-hot failure also has a generative-family reading (discrete diffusion suits categorical targets (Austin et al., 2021; Lou et al., 2024)), tested by two controls (Appendix N). The converse route, continuous encodings of discrete data for Gaussian diffusion (Chen et al., 2023; Han et al., 2023; Hoogeboom et al., 2021), is the family of both targets, one-hot its position-free member, phase field its graded one (Sec. 3.2).

Generative models for physiological signals and diffusion. Diffusion models serve ECG/EEG generation and super-resolution (Liu et al., 2026; Wang et al., 2026); our three-layer protocol inherits their layered-evaluation idea (Wang et al., 2026; Liu et al., 2026). ECG synthesis has a longer line (McSharry et al., 2003; Lopez Alcaraz & Strodthoff, 2023); diffusion has entered delineation itself for contaminated signals (Zhu et al., 2024), with segmentation accuracy as the object; we ask which target the prior should generate and whether it survives averaging, scored by sequence legality on a measured-noise grid. Closest to our setting: diffusion and flow matching generate masks or their distance fields in images (Amit et al., 2021; Bogensperger et al., 2023; 2025; Wu et al., 2024; Zhang et al., 2025), segment from diffusion representations (Baranchuk et al., 2022), and denoise per-frame labels in action segmentation (Liu et al., 2023; 2025), the nearest temporal instance; DiffAct reads out a single deterministic trajectory, the exact $K{=}1$ operating point where one-hot proves fragile here, with best-of-K only a diagnostic ceiling. Three differences delimit the contribution: the constraint is a *temporal grammar* (legality, not mask overlap), the operating claim is *a single draw at rules' level* (not only selection), and the object of study is the *averagability of the target* (one-hot means collapse at the decision-boundary readout, field means do not within the measured regime; Sec. 4.4, Appendices P and Q). We generate *structure*, not signal: diffusion carries a structural prior; the contribution is *what target it should generate*.

3 METHOD

3.1 PROBLEM SETTING AND SEQUENCE LEGALITY

The input is a single-lead signal $x \in \mathbb{R}^S$ (500 Hz), and the delineator outputs a mask $m \in \{0,1,2,3\}^S$ (background/P/QRS/T). A *run* is a maximal connected segment of same-class samples; sorting runs by onset yields the class sequence $\sigma(m)$ (same-class runs separated by background stay separate). The structural template is the regular grammar

$$(P?\ \ QRS\ \ T)^+, \tag{1}$$

and the implemented legality check, used for every legality number in this paper, is its transition relation: every adjacent-run transition must belong to $\{P{\to}QRS,\ QRS{\to}T,\ T{\to}P,\ T{\to}QRS\}$. The check is deliberately more permissive than the full language of Eq. 1: evaluation windows

truncate beats at either end, so the strict language's boundary conditions are omitted (strict: 41% of ground truth parses vs. 95.9%; the census of the ground truth's own illegal 4.1%: Appendix A). *Legality* is the fraction of legal lead-samples, the evaluation unit throughout (one windowed signal channel); records serve as the resampling unit for all intervals (Sec. 4.1). P is optional in the grammar, QRS and T mandatory: this reflects physiology (P waves may be indiscernible under certain rhythms or leads (Hindricks et al., 2021; Moody & Mark, 1983)) and marks the metric's shortest gaming path, deleting all P waves, for which the wave-deletion audit (Sec. 4.1) is the counter-check. Sequences the grammar cannot express, such as the non-conducted P of second-degree block, are censused in Appendix A (limitation: Appendix K). Illegality has two sources: *fragmentation* (same-class runs split and interleave with other classes) and *misordering* (illegal transitions, e.g., $QRS{\to}P$, $P{\to}T$).

Legality preconditions interval measurability (PR needs the $(P_{\mathrm{on}}, QRS_{\mathrm{on}})$ pair complete and consistently attributed, QT likewise $(QRS_{\mathrm{on}}, T_{\mathrm{off}})$); the failure is local but the verdict global, and a deployment pipeline without localization or repair can only reject the record wholesale.

3.2 The phase-field family

Encoding. We assign one channel (or channel pair) per wave class and parameterize the value by *within-wave phase* (the name refers to this phase, not to physics interface fields). A run $[o, d)$ (onset o, duration d) takes values

$$z_c[o{+}i] \;=\; \sin\big(\pi\,\phi_i\big)\, e^{i\pi\phi_i}, \qquad \phi_i = i/d \in [0, 1), \tag{2}$$

the *complex phase field*: the amplitude $\sin(\pi\phi)$ rises to 1 at the run center and falls back to 0, the angle $\pi\phi$ encodes the within-wave position, and the background is 0. The family of continuous targets compared in Sec. 4.4 arises by restricting this field to its coordinates: the **pure phase field** $e^{i\pi\phi}$ (the angle alone, $|z| \equiv 1$; between waves the angle interpolates along the lower half-circle, presence becoming the sine sign), the **complex field** z itself (amplitude and angle jointly, 6 real channels), and the **amplitude field** $\sin(\pi\phi)$ (the modulus alone, 3 channels). We adopt the simplest member, the amplitude field, as the main configuration and abbreviate it *the field*; *phase field* names the family throughout. Sec. 4.4 shows the modulus is the sufficient coordinate. Three properties make the amplitude field a structural target:

1. C^0 **continuity**: values never jump; boundaries are crossed through zero.
2. **Invertibility**: decoding is closed-form (below); the mask↔field round trip has boundary error of at most one sample (480 LUDB validation lead-samples).
3. **All three structural elements live in the target**: presence (non-zero channel), boundary (zero-crossing position), and order (class is carried by the channel; decoding needs no sorting).

The default one-hot target, by contrast, carries no within-wave position: a sample at the wave center and one at the edge receive the same value (Fig. 4, Appendix L); the amplitude field grades them instead.

Decoding. Connected components with channel value $a > \tau$ ($\tau = 0.15$; the support covers about 90% of the run length) form a run's *support*; from the analytic form of $\sin(\pi\phi)$, outward extrapolation by $\phi_\tau \cdot d$ at both ends ($\phi_\tau = \arcsin(\tau)/\pi$) recovers the true boundary (Fig. 5b, Appendix L). Classes are decoded independently in a fixed order (P, QRS, T; on rare τ-support overlaps the later class wins), background being the untouched initial value; training-target supports are disjoint by construction. The decoder is thresholding plus analytic extrapolation: *no merging, no reordering, no grammar rules*; all structural repair duty moves from the decoder into the representation. A decoder-symmetry control shows the decoder carries none of the selection-tier gap, its readings unchanged under matched decoders (Appendix P); on the mean tier, the one-hot collapse is specific to the decision-boundary readout, scoping the averagability claim of Sec. 4.4.

3.3 Conditional diffusion over fields and aggregation operators

Diffusion prior. A U-Net denoiser U_θ (the same capacity as the discriminative anchor, 1.77M parameters) is trained over the field space as a conditional diffusion model: the training objective is the

x_0-prediction mean squared error $\|U_\theta(x_t, t, \tilde{x}) - z_0\|^2$, where x_t is the diffusion-noised ground-truth target z_0 (the field, or the one-hot channels themselves; the focal-on-logits convention is a control in Appendix N) and $\tilde{x}$ is the corrupted input signal; sampling uses 20 steps of denoising diffusion implicit models (DDIM) (Ho et al., 2020; Song et al., 2021).

In-model noise corruption. During training, each batch applies measured noise to the input signal with 75% probability (baseline wander, electrode motion, and muscle artifact from the MIT-BIH Noise Stress Test Database (NSTDB), plus Gaussian; SNR $\sim \mathrm{U}[0, 12]$ dB (Moody et al., 1984)) and stays clean with 25% probability. Corruption is necessary: without it (same capacity, same protocol) a focal-convention one-hot diffusion control's worst-cell legality collapses to 0.334 (Appendix I).

Aggregation operators. Diffusion sampling is stochastic; one inference call produces K independent candidate fields, aggregated in two ways:

- **legality_select** (main operating point): decode each candidate and take the one with the fewest illegal transitions, ties broken by candidate order. No ground truth, no repair: the grammar only *scores* candidates, never edits them; the field decoder is grammar-free (tie handling: Appendix A).
- **mean_field**: average the K fields point-wise and decode once. Sec. 4.4 shows that one-hot collapses while the field stays robust under this operator, within the measured regime (decode- and regime-conditional: Appendices P and Q).

4 Experiments

4.1 Evaluation protocol and data

Protocol. Any single metric for "structured output under noise" can be gamed, so we evaluate in three layers, one question each: *structure* (legality, worst of 12 conditions; the **primary metric**), *usability* (GT-anchored audits; validation axis), and *precision* (F1, intervals, tolerance scans; quality control). Legality is primary because it preconditions interval measurability (Sec. 3.1) and *requires no ground truth*, scaling to unlabeled signals (real AF: Appendix F); the beat-level full audit is the validation axis (Sec. 4.5). Deleting all P waves would leave legality intact; the wave-deletion audit counts that path (Sec. 4.3).

Data and noise grid. **Training/validation**: LUDB (Kalyakulina et al., 2020), 200 records split 160/40 into single-lead samples (protocol: Appendix A). **External test**: QTDB with revised labels (Jimenez-Perez, 2021), originally Laguna et al. (1997); 103 records / 1,676 lead-samples (label quality: Appendix A); training and test are disjoint and cross-database. All signals and noise sources are resampled to a common 500 Hz (native rates and operators: Appendix A). **Noise grid**: 12 conditions = clean + gauss@5/10 dB + bw/em/ma@$\{0, 6, 12\}$ dB (measured NSTDB noises), fixed shared seeds (paired comparison); *all variants receive only the noisy signal itself, never the noise type or level*. The primary metric is reported at the *worst cell* (a variant's minimum legality over the 12 conditions; em@0 is the harshest).

Usability audits. A beat (a ground-truth QRS run) receives one of six structure-only verdicts (*full*/*partial*/*frag*/*miss*) over two spans, the *measurement track* and the *cycle track* (definitions, attribution, and basis: Appendix G). The *wave-deletion audit* counts *deleted* ground-truth and *spurious* predicted runs per wave class: whether legality is bought by deletion.

Operating points and statistics. Learned variants report $K{=}8$ legality select; the anchor and the rules variant are deterministic single forwards; diffusion variants additionally report $K{=}1$ and mean aggregation. $K{=}8$ is fixed before test evaluation, the K-scan reporting it as a budget sensitivity (Sec. 4.4). F1 is the grid mean of per-point F1 at 150 ms over the six fiducials; intervals are median absolute errors on ground-truth-paired beats (pairing: Appendix H); *full* is the per-record mean. Differences report 10,000-resample paired bootstrap CIs with records as resampling units (the 1,676 lead-samples cluster into 103 records); significance is a CI excluding zero. These are per-comparison intervals without multiplicity correction; near-zero lower bounds are read as point estimates, not

Table 1: Main results (external QTDB test, revised labels, 12-condition grid; diffusion variants at $K{=}8$ select, all with the same x_0-MSE objective; $K{=}1$/mean readings: Sec. 4.4).

Variant (same backbone)	F1@150 ms	Leg. mean	Leg. worst
Discriminative anchor	0.865	0.717	0.544
+ rule post-processing	0.870	0.969	0.863
+ one-hot diffusion	<u>0.875</u>	<u>0.990</u>	<u>0.980</u>
+ **phase-field diffusion** (ours)	**0.880**	**0.999**	**0.994**

certified orderings (flagged where they arise). Column-best is bold, second-best underlined (lowest in count/error columns).

4.2 Main results: four variants on the same anchor

Same-anchor isolation. All variants share one backbone, the capacity-matched discriminative anchor with in-model corruption: rules add post-processing to it, the diffusion variants retrain it with a diffusion objective, so both face the same detector's output distribution (Table 1).

The four-variant chain and its mechanism (em@0, every variant's worst cell): bare anchor 0.544 $\rightarrow$ rules 0.863 $\rightarrow$ one-hot diffusion 0.980 $\rightarrow$ field diffusion 0.994. Rules recover $+31.9$ [28.69, 35.07]; the one-hot diffusion $+43.62$ [39.00, 47.96] over the anchor ($+11.75$ [9.21, 14.43] over rules), the field's CIs below; the one-hot advantage over rules is certified in the heavy-noise cells and crosses zero under clean and light noise (census in Appendix B).

Against rules, all six audited conditions exclude zero. The field diffusion's paired-bootstrap legality gains run from em@0 $+13.13$ $[10.20, 16.19]$ down to clean $+1.01$, all six audited cells excluding zero; the joint worst-cell CI gives field$-$rules $+13.13$ $[10.15, 16.22]$, matching the em@0 pairing because every variant's worst cell sits at em@0 (construction: Appendix B). Over the bare anchor, em@0 is $+44.99$ $[40.07, 49.72]$, and over the strongest external baseline (a 3-filter ECG-UNet, 1.84M) $+50.60$ $[46.11, 55.15]$. The external baseline *shares our noise recipe* and the same early-stopping budget (Appendix A), yet its worst cell is 0.488, in the same fragility tier as the bare anchor's 0.544 (no ordering is read), with grid-mean F1 0.856. Learned variants are early-stopped on validation F1, nearly blind to legality (Sec. 4.1): the gains are not a model-selection artifact. The combination cell (the same assembler stacked on the diffusion outputs; Appendix O) is a bounded null on the field ($+0.18$ pp worst cell, CI crossing zero; on the matched one-hot, $+0.4$ pp at em@0, likewise crossing zero): repair has moved into the representation; what the assembler still buys is inter-beat curation (Sec. 4.5).

The selection-tier representation step is real but narrow. Replacing the target (one-hot $\rightarrow$ field) at the same loss adds $+1.37$ $[0.48, 2.42]$ at em@0 (ma@0 $+1.01$ $[0.41, 1.73]$; light noise $+0.6$–1.0, several touching zero; joint worst-cell $+1.37$ $[0.49, 2.34]$ (seed-42 protocol); under the focal convention $+6.44$ $[4.55, 8.48]$ at em@0 and $+6.56$ $[4.98, 8.75]$ joint; the per-seed em@0 steps span $+0.06$ to $+1.55$, seed 43's interval crossing zero, the three-seed merged step $+0.99$ $[0.39, 1.75]$ excluding zero; Appendices B, N, P); the load-bearing representation gains are the selection-free ones of Sec. 4.4. F1 moves in step ($+2.23$ $[1.46, 3.05]$ generative-head, per-record; Appendix J; the representation step is $+0.5$ pooled, a point estimate, $+0.80$ $[0.12, 1.47]$ per-record under the focal convention): legality gains cost no precision, each basis compared within itself.

Dominance and robustness. The field diffusion occupies the top legality tier (worst cell 0.994, on par with a detection-cascade variant, Appendix D); no variant reverses it in any condition (Fig. 2; per-cell values in Appendix B). Across three training seeds the worst cell stays at 0.9928–0.9952, with neither the one-hot band (0.979–0.992) nor the rules band (0.863–0.913) overlapping it; the one-hot margin is thin (0.08 pp), so the ordering rests on the paired intervals above, not the band gap (all bands incl. focal 0.928–0.968: Appendices A, N). The GT-free metric also carries evaluation to unlabeled real signals: on real AF episodes from a disjoint database (a P-less rhythm) legality is ours 0.88 $\rightarrow$ 0.86 vs. rules 0.81 $\rightarrow$ 0.70 (clean $\rightarrow$ em@6, both CIs excluding zero; a lower absolute tier,

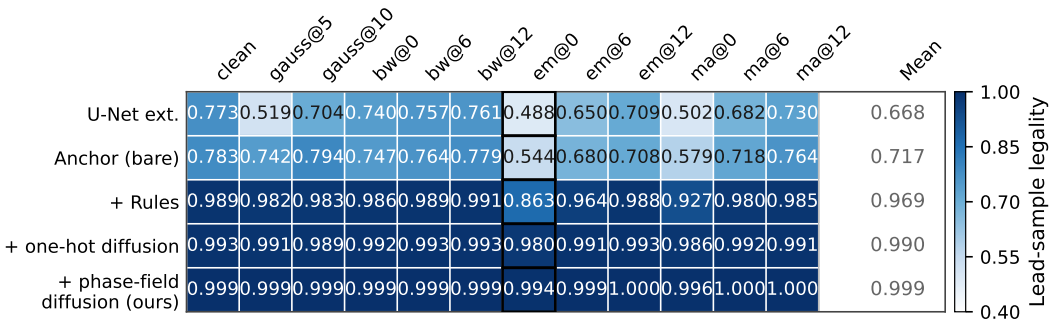


Figure 2: Legality over the 12-condition grid (variant × condition; column labels such as gauss@5 abbreviate gauss@5 dB; color encodes legality). Rows, top to bottom: the external U-Net baseline, the bare anchor, + rules, + one-hot diffusion, and + phase-field diffusion (ours); the gray Mean column averages legality over the 12 conditions. Worst cells are boxed, every variant's at em@0: the rules variant collapses in the heavy-noise columns while the two diffusion variants stay ≥ 0.98 there.

revisited in Appendix K), with companion audits ruling out hallucinated P and unsupported QRS (Appendix F).

4.3 Generation vs. rules: the wave-deletion audit

Three readings. **First, the principle boundary of rules is quantified**: after rule repair, more ground-truth P waves are deleted than by the bare detector, and T deletions nearly double (Table 2); rules trade fewer spurious runs for more deletions, the quantitative form of "cannot create from nothing." The residual shows where it bites: among the 230 illegal lead-samples rules still leave, transitions are dominated by $qrs{\to}p/qrs{\to}qrs$; once T is missed, reassembly cannot repair what follows (Appendix A). **Second, generation supplies exactly what is missing, for either representation**: both diffusion variants delete fewer ground-truth P waves than either non-generative variant; the field deletes 13.7% fewer than rules (-45.4% on T) at the highest repaired P recall (0.754; one-hot 1,628, recall 0.735) and carries fewer spurious runs than the one-hot on every class (rules stay leaner still on spurious P, 826 vs. our 980): wave presence is not sequence legality (legality 0.994 vs. 0.980, frag 0.027 vs. 0.034, Table 3), so the deletion audit alone does not pick the winner. **Third, rules are better on QRS**: deletions 98 vs. 161, recall 0.986 vs. 0.976; two-thirds of both counts are window-edge truncation artifacts and the interior difference is small (census: Appendix A). Hence "fewer P/T misses," not "fewer misses." Fig. 6 shows a representative record (Appendix M).

4.4 Representation and sampling: the K-scan

All continuous fields beat one-hot; the simplest suffices at every operating point. Fig. 3 traces both aggregation tracks at em@0 for four target representations plus two one-hot controls: the three continuous fields cluster (single draws 0.867–0.888, means within 0.76–0.87 at every K), while the one-hot starts at 0.794 and its mean collapses completely at $K{=}2$ (0.000 in all twelve conditions; mean@16 recovers only to 0.229). Both controls sit lower still (discrete diffusion below the Gaussian one-hot at every selection K, $K{=}16$ ceiling 0.829 $<$ rules' 0.863, mean collapse alike; focal single draw 0.349; Appendix N): collapse follows the target, not the sampler family or the loss, within

Table 2: Wave-deletion audit at em@0: ground-truth runs deleted / spurious predicted runs, and wave-level recall.

Variant	P deleted	T deleted	QRS deleted	P recall
Discriminative anchor (bare)	1,686 / 1,126	**494** / 2,004	**92** / 1,635	0.726
+ rule post-processing	1,751 / **826**	906 / **1,093**	98 / 1,265	0.715
+ one-hot diffusion ($K{=}8$)	1,628 / 1,070	542 / 1,283	227 / 987	0.735
+ phase-field diffusion (ours)	**1,511** / 980	495 / 1,215	161 / **973**	**0.754**

Table 3: GT-anchored beat audit at em@0 (pooled; full/full_cyc: measurement/cycle track; per-record values differ by $\leq$1 pp; K=1 rows, the baseline row, and per-record CIs: Appendix G).

Variant	full	full_cyc	qt_usable	frag	frag_cyc
Discriminative anchor	0.694	0.574	0.863	0.107	0.250
+ rule post-processing	0.738	**0.665**	0.927	0.036	**0.123**
+ one-hot diffusion (K=8)	0.725	0.617	0.924	0.034	0.162
+ phase-field diffusion (K=8, ours)	**0.763**	0.644	**0.938**	**0.027**	0.161

the tested budget at the decision-boundary readout; the decode axis itself is swept in Appendix P. Among the family members of Sec. 3.2, the modulus is the sufficient coordinate (best worst cell 0.994 vs. 0.986/0.992, best F1 0.880 vs. 0.861/0.872): the angle buys a rounding-level margin and costs precision. *Averagability and single-draw robustness are properties shared by every continuous vector target tested here*; the one-hot mean collapse is, in addition, decode-conditional (total at the decision-boundary readout, partially rescuable by off-boundary thresholds; Appendix P), while boundary precision and the worst cell are still best under the amplitude implementation.

At K=1 the field's 0.888 clears rules' 0.863 at the seed-42 pairing ($+2.57$ $[0.37, 4.82]$), though the seed bands overlap (0.876–0.896 vs. 0.863–0.913): the seed-robust statement is the representation step, positive at every seed ($+3.2$ to $+11.4$ pp; $+9.49$ $[6.74, 12.24]$ at seed 42; focal: $+53.9$, Appendix N), while the one-hot sits below rules at every seed (band 0.762–0.864; -6.92 $[-9.66, -4.37]$). The edge is a heavy-noise phenomenon: away from the two harshest cells the single draw trails rules by 1–3 pp (clean -1.61 $[-2.56, -0.77]$), the K=1 face of Sec. 4.5's clean conservatism, erased by selection. At K=8 the field's 0.994 stands above the widest rules seed (0.913) by $\geq$8 pp, while the one-hot's 0.980 clears it by 6.7 pp yet leaves a narrow step to the field (Sec. 4.2). *The field's selection gain is* $+10.6$*, the representation step* $+1.4$ (0.994 vs. 0.980 at K=8), the target-side difference at the same operating point, bought by the representation, not by samples. The contrast is in the candidate pool (Fig. 3a): at K=2 the field holds a legal candidate for 95.8% of lead-samples (one-hot 90.5%; focal convention 56.9%, Appendix N). The rules-level single draw costs $\approx$18$\times$ rules, the K=8 tier $\approx$145$\times$, yet both process a 4 s window within real time on one consumer GPU ($0.02\times$ and $0.19\times$, batch-1 throughput, not per-beat monitoring latency; Appendix A).

Attribution and scope. The mechanism is the decoding geometry (Fig. 5a, Appendix L): one-hot's per-point argmax is not closed under averaging, as misaligned samples average to a plateau whose argmax flips near the boundary, fragmenting into illegal sequences; the field tolerates averaging instead, its mean staying legal at every measured K (0.76–0.87), an empirical robustness, not a closure guarantee (mean-decode legality conditions: Appendix L). A cross-domain probe extends the claim: on heart-sound segmentation, the single-draw and mean@2 stability conclusions replicate (Appendix Q). The anti-diagonal swap (Appendix C) and the cascade control (Appendix D) pin the effect to the target side relative to the network, conditioning and head leaving it untouched; on the

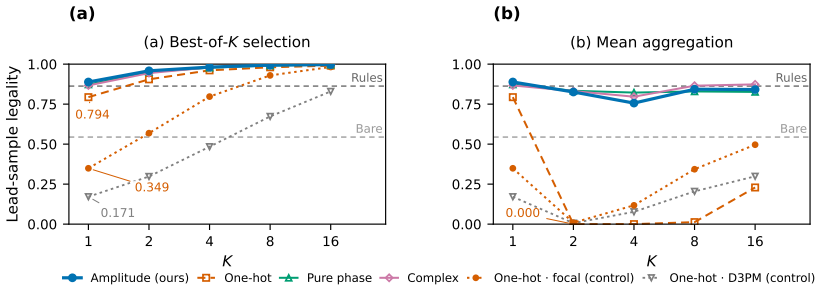


Figure 3: K-scan at em@0 for four target representations (the one-hot trained with the same x_0-MSE loss as the fields) plus two one-hot controls, $K \in \{1, 2, 4, 8, 16\}$ (legality; dashed: rules 0.863 and the bare anchor 0.544; dotted vermillion: a focal-on-logits control; gray dotted: absorbing-state discrete diffusion; both Appendix N). (a) Best-of-K selection, the any-k ceiling by construction; (b) mean aggregation. Readings: Sec. 4.4.

decode axis the collapse is specific to the decision-boundary readout, and the field averages legally with no decode tuning (Appendix P). The absorbing-state control's early-reveal commitment amplifies the same fragility (Appendix N; other kernels untested). The scope: a deterministic regressor onto the field (same backbone, target, corruption; Appendix J) keeps a third of the diffusion gain and attains the best F1 (0.885), yet saturates below rules (0.682): under heavy noise the posterior over fields is multimodal and the conditional mean smears instead of committing (Eikema & Aziz, 2020); missed waves stay missed, and only sampling synthesizes them. The representation decides the robustness of *generative* structured prediction, and supplies precision and a legality floor even deterministically.

4.5 Validation and precision layers

Measurement track: structural gain converts into usability. Under noise, our full is significantly above rules (em@0 $+2.63$) and above the matched one-hot in every noisy condition (clean a tie; per-condition: Appendix G), with the lowest frag (Table 3) and the mean operator intact where one-hot's collapses (mean@2 0.826 vs. 0.000 at the decision-boundary readout; Appendix P). Two qualifications. *(i) Clean reversal*: under clean, rules have higher full (-1.00, a point estimate; Sec. 4.1) while legality still favors us ($+1.01$ $[0.47, 1.64]$): beat-level completeness slightly trails, structure does not; the τ scan locates it as "field-did-not-paint" misses, one-hot behaving alike (Appendix E). *(ii) Rules win the cycle track everywhere* (full_cyc em@0 -2.95, clean -8.32): the gap comes from TP-segment ghost waves (ours 1.99/record, fewest among learned); rules' assembly discards unattributable runs (1.73/record); its legality deficit is missing waves, not extra ones. Inter-beat looseness is the diffusion head's improvement point (cycle-aware sampling; bounds: Appendix O).

Intervals trail rules by at most 4 ms (PR $+0$–4, QT $+2$–4, QRS $\leq +2$; PR ties at em@0); *coverage is disclosed per condition*: under noise ours pairs the most beats, under clean rules pair slightly more (per-condition medians and counts: Appendix H). Margins of this size sit within the repeatability of manual interval measurement itself, so the structural gains above come with no interval cost a clinical reader would notice. Cross-domain error is roughly twice in-distribution (QTDB medians 12–28 ms vs. LUDB val 8–18 ms), and this holds for both variants with in-distribution readings (Table 6): a database-level difference, not a property of our method.

Precision bundle. F1 0.880 is the best of the four-variant chain (rules 0.870). The tolerance scan preserves the ordering, and a channel-held-out grid (the NSTDB test noise drawn from the channel unseen during training) changes mean legality by -0.05 pp (both in Appendix E).

5 Conclusion

Limitations. Two rank above the rest: the cycle track is below rules in all conditions (1.99 TP ghost waves/record; the reachable gain is bounded in Appendix O), and clean-condition full slightly trails rules (-1.00; located by the τ scan in Sec. 4.5). Seven further limitations are detailed in Appendix K.

Conclusion. We elevated structural failure under noise to a first-class problem, measured it with sequence legality under a three-layer protocol, replaced rule repair with generation ($+45.0$ pp, 13.7% fewer missed P waves), and moved structure into the representation: a phase field whose single draw sits at rules' level at the harshest cell, above the matched one-hot at every seed, whose averaging never collapses at any tested decode threshold, and whose selection reaches 0.994 above the one-hot's 0.980. **Under the harshest noise, generation replaces rule repair; structure steers generation.** A heart-sound probe replicates the single-draw and mean@2 stability conclusions (Appendix Q); the representation-level lesson may extend to other grammar- or topology-constrained structured-prediction tasks.

Reproducibility statement

All experiments use public datasets (LUDB, QTDB with revised labels, MIT-BIH NSTDB and AFDB via PhysioNet, and PhysioNet/CinC Challenge 2016 for the heart-sound probe of Appendix Q). Every number in this paper traces to a versioned artifact under `results/` produced by scripts in `scripts/`; noise seeds, training seeds, and the paired bootstrap protocol are fixed in code.

Training a single variant takes 6–27 minutes on one consumer GPU (RTX 3080 Ti Laptop), and the experiments reported here total well under 10 GPU-hours. We will release code, configurations, and evaluation scripts upon publication.

ETHICS STATEMENT

This work uses only publicly available, de-identified research datasets and involves no personal health information. The method is a signal-processing research contribution; clinical use requires further validation, and this work is research code, not a medical device. Four deployment-relevant cautions. First, a generative delineator can synthesize plausible-looking waves that the signal does not support: our audits bound this risk on the tested corpora (fewer ground-truth P waves missed than rules, no per-beat P hallucination on AF segments, QRS anchored to annotated R peaks; Sec. 4.3 and Appendix F), but hallucination under distributions beyond our grid remains possible and would propagate into interval measurements silently. Second, interval errors are of clinical scale only in context: the compared variants differ by at most 4 ms in QT/PR medians, below the ˜5–10 ms repeatability of manual QT measurement, while the absolute cross-domain errors (QT medians 12–28 ms) remain of clinical scale. Third, the operating claims are bounded by the evaluation surface: single-lead 500 Hz records, LUDB training, QTDB/AFDB external tests, and a noise grid of three measured artifacts plus Gaussian; rhythms, leads, and artifacts outside this surface are unverified. Fourth, the grammar itself can reward deleting real waves: on rhythms it cannot express (non-conducted P waves in second-degree block, where the census of Appendix A shows legality is achieved by dropping the extra P), a grammar-compliant output is structurally lossy for exactly the conduction-disorder population whose PR measurement matters most, and the affected intervals are silently dropped rather than flagged.

LLM USE DISCLOSURE

We used large-language-model assistants (ZCode, GLM) as writing and coding aids: manuscript editing and compression, LaTeX plumbing, and boilerplate for evaluation scripts. All technical content (the method design, experiment design, all numbers, and all conclusions) is the authors' own and was verified against versioned artifacts under `results/`; every quantitative claim in this paper traces to a script-generated JSON artifact, and the statistical scripts were re-audited line by line during revision. No LLM was used to generate experimental results, and no LLM-produced text is presented as externally sourced fact.

REFERENCES

Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021.

Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2858–2866, 2017. doi: 10.1109/CVPR.2017.305. arXiv:1611.08303.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

David Belanger and Andrew McCallum. Structured prediction energy networks. In *International Conference on Machine Learning*, volume 48, pp. 983–992, 2016.

Lea Bogensperger, Dominik Narnhofer, Filip Ilic, and Thomas Pock. Score-based generative models for medical image segmentation using signed distance functions. In *German Conference on Pattern Recognition (GCPR)*, 2023. arXiv:2303.05966.

Lea Bogensperger, Dominik Narnhofer, Andreas Falk, Konrad Schindler, and Thomas Pock. FlowSDF: Flow matching for medical image segmentation using distance transforms. *International Journal of Computer Vision*, 133(7):4864–4876, 2025. arXiv:2405.18087.

Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *International Conference on Learning Representations (ICLR)*, 2023.

Daniel Deutsch, Shyam Upadhyay, and Dan Roth. A general-purpose algorithm for constrained sequential inference. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019.

Luigi Y. Di Marco and Lorenzo Chiari. A wavelet-based ECG delineation algorithm for 32-bit integer online processing. *BioMedical Engineering OnLine*, 10:23, 2011. doi: 10.1186/1475-925X-10-23.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1801–1810, 2019. arXiv:1904.07846.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020.

Ary L. Goldberger, Luís A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, C.-K. Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. SSD-LM: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Gerhard Hindricks, Tatjana Potpara, Nikolaos Dagres, Elena Arbelo, Jeroen J. Bax, Carina Blomström-Lundqvist, et al. 2020 ESC guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the european association for cardio-thoracic surgery (EACTS). *European Heart Journal*, 42(5):373–498, 2021. doi: 10.1093/eurheartj/ehaa612.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. arXiv:2006.11239.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Guillermo Jimenez-Perez. QTDB annotations. figshare, revised delineations for the QT database, 2021.

Guillermo Jimenez-Perez, Alejandro Alcaine, and Oscar Camara. Delineation of the electrocardiogram with a mixed-quality-annotations dataset using convolutional neural networks. *Scientific Reports*, 11:863, 2021. doi: 10.1038/s41598-020-79512-7.

Guillermo Jimenez-Perez, Juan Acosta, Alejandro Alcaine, and Oscar Camara. Generalising electrocardiogram detection and delineation: Training convolutional neural networks with synthetic data augmentation. *Frontiers in Cardiovascular Medicine*, 11:1341786, 2024. doi: 10.3389/fcvm.2024.1341786.

Alena I. Kalyakulina, Igor I. Yusipov, Viktor A. Moskalenko, Alexander V. Nikolskiy, Konstantin A. Kosonogov, Grigory V. Osipov, Nikolai Yu. Zolotykh, and Mikhail V. Ivanchenko. LUDB: A new open-access validation tool for electrocardiogram delineation algorithms. *IEEE Access*, 8: 186181–186190, 2020. dataset hosted on PhysioNet, `https://physionet.org/content/ludb/1.0.1/`.

Paul Kligfield, Leonard S. Gettes, James J. Bailey, Rory Childers, Barbara J. Deal, E. William Hancock, Gerard van Herpen, Jan A. Kors, Peter Macfarlane, David M. Mirvis, Olle Pahlm, Pentti Rautaharju, and Galen S. Wagner. Recommendations for the standardization and interpretation of the electrocardiogram: Part I: The electrocardiogram and its technology. *Circulation*, 115(10): 1306–1324, 2007. doi: 10.1161/CIRCULATIONAHA.106.180200.

Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of HLT-NAACL*, 2004.

Pablo Laguna, Roger G. Mark, Ary L. Goldberger, and George B. Moody. A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG. In *Computers in Cardiology 1997*, pp. 673–676, Lund, Sweden, 1997. Data hosted on PhysioNet: `https://physionet.org/content/qtdb/1.0.0/`.

Chengyu Liu, David Springer, Qiao Li, Benjamin Moody, Ricardo Abad Juan, Francisco J Chorro, Francisco Castells, José Millet Roig, Ikaro Silva, Alistair E W Johnson, Zeeshan Syed, Samuel E Schmidt, Chrysa D Papadaniil, Leontios Hadjileontiadis, Hosein Naseri, Ali Moukadem, Alain Dieterlen, Christian Brandt, Hong Tang, Maryam Samieinasab, Mohammad Reza Samieinasab, Reza Sameni, Roger G Mark, and Gari D Clifford. An open access database for the evaluation of heart sound algorithms. *Physiological Measurement*, 37(12):2181–2213, 2016.

Daochang Liu, Qiyue Li, AnhDung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion action segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. DiffAct++: Diffusion action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1644–1659, 2025. doi: 10.1109/TPAMI.2024.3509434.

Hongjun Liu, Leyu Zhou, Zijianghao Yang, and Chao Yao. Step-aware residual-guided diffusion for EEG spatial super-resolution. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.19166.

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. doi: 10.1016/j.compbiomed.2023.107115.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, 2024. arXiv:2310.16834.

Jun Ma, Zhan Wei, Yiwen Zhang, Yixin Wang, Rongfei Lv, Cheng Zhu, Gaoxiang Chen, Jianan Liu, Chao Peng, Lei Wang, Yunpeng Wang, and Jianan Chen. How distance transform maps boost segmentation CNNs: An empirical study. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning (MIDL)*, volume 121 of *Proceedings of Machine Learning Research*, pp. 479–492, 2020.

Juan Pablo Martínez, Rute Almeida, Salvador Olmos, Ana Paula Rocha, and Pablo Laguna. A wavelet-based ECG delineator: Evaluation on standard databases. *IEEE Transactions on Biomedical Engineering*, 51(4):570–581, 2004. doi: 10.1109/TBME.2003.821031.

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. doi: 10.1109/TBME.2003.808805.

George B. Moody and Roger G. Mark. A new method for detecting atrial fibrillation using R-R intervals. In *Computers in Cardiology*, volume 10, pp. 227–230, 1983. MIT-BIH Atrial Fibrillation Database, `https://physionet.org/content/afdb/1.0.0/`.

George B. Moody, Warren E. Muldrow, and Roger G. Mark. A noise stress test for arrhythmia detectors. In *Computers in Cardiology 1984*, volume 11, pp. 381–384, 1984. Data hosted on PhysioNet: `https://physionet.org/content/nstdb/1.0.0/`.

Viktor Moskalenko, Nikolai Zolotykh, and Grigory Osipov. Deep learning for ECG segmentation. In *Advances in Neural Computation, Machine Learning, and Cognitive Research III*, volume 856 of *Studies in Computational Intelligence*, pp. 246–254. Springer, 2020. doi: 10.1007/978-3-030-30425-6_29. arXiv:2001.04689.

Jiapu Pan and Willis J. Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985. doi: 10.1109/TBME.1985.325532.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015)*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2018)*, volume 11071 of *Lecture Notes in Computer Science*, pp. 265–273. Springer, 2018. doi: 10.1007/978-3-030-00934-2_30.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. arXiv:2010.02502.

David B. Springer, Lionel Tarassenko, and Gari D. Clifford. Logistic regression-HSMM-based heart sound segmentation. *IEEE Transactions on Biomedical Engineering*, 63:822–832, 2016.

Tibor Stracina, Marina Ronzhina, Richard Redina, and Marie Novakova. Golden standard or obsolete method? review of ECG applications in clinical and experimental context. *Frontiers in Physiology*, 13:867033, 2022. doi: 10.3389/fphys.2022.867033.

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

Longsheng Wei and Yuyang Ye. Fine-grained action segmentation based on complementary frame-level classification model and action-wise regression model. *Displays*, 74:102212, 2022. doi: 10.1016/j.displa.2022.102212.

Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, volume 227, pp. 1623–1639, 2024. PMLR.

Hantao Zhang, Yuhe Liu, Jiancheng Yang, Weidong Guo, Xinyuan Wang, and Pascal Fua. Diffatlas: Genai-fying atlas segmentation via image-mask diffusion. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2025)*, 2025.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1529–1537, 2015. doi: 10.1109/ICCV.2015.179.

Baofeng Zhu, Wanjun Cheng, Lihua Sun, Yingying Feng, Xia Zhang, and Jiren Liu. Diffusion based comprehensive approach for highly contaminated electrocardiogram segmentation. *Biomedical Signal Processing and Control*, 97:106693, 2024. doi: 10.1016/j.bspc.2024.106693.

A Implementation and hyperparameters

Legality-check scope. The transition check of Sec. 3.1 deliberately omits the boundary conditions that the strict language of Eq. (1) imposes, $\sigma_1 \in \{P, QRS\}$ and $\sigma_{\text{last}} = T$ ($\sigma(m)$ is the onset-sorted run sequence; same-class runs separated by background stay separate): evaluation windows cut at either end, so a window may legitimately begin with a residual T or end mid-beat (a trailing P or QRS). Under the strict grammar only 41% of ground-truth segments parse (window truncation), while ground truth itself reaches 95.9% under the transition check.

Ground-truth illegality census (the 95.9% ceiling). Ground truth itself fails the transition check on 68 of the 1,676 lead-samples (4.1%), concentrated in 15 records, with three distinct causes. (i) *Unannotated T in AF*: sel35 (16 lead-samples) is flagged in the QTDB documentation as an atrial-fibrillation excerpt, and all of its 46 illegal $qrs{\to}qrs$ pairs (of the corpus-wide 50) span a median 840 ms QRS-to-QRS gap: beats whose T the annotators did not delimit. (ii) *Skipped QRS annotations*: 38 $p{\to}t$ and 28 $p{\to}p$ pairs, dominated by sel46 and sel231; cross-checked against the original QTDB manual beat annotations (the annotator-1 `.q1c` beats, which play no role in building the corrected masks): 36 of the 38 $p{\to}t$ gaps contain an annotated beat (the corrected-label annotator skipped it), and *none* of the 28 $p{\to}p$ gaps does: genuinely non-conducted P waves, a second-degree-block sequence the grammar cannot express by construction. (iii) *Split annotations*: 16 $t{\to}t$ pairs 4–8 samples apart (largest share: sel45). The 16 windows that carry no ground-truth QRS at all are where cause (ii) concentrates: they account for 36 of the 38 $p{\to}t$ pairs and 2 of the 28 $p{\to}p$ pairs: whole windows whose beats were never annotated. What the variants do on these 68 lead-samples closes the legality tier above the GT rate: at clean the field's $K{=}8$ output is legal on 68/68 (em@0 67/68), the one-hot's on 67/68 (em@0 65/68), rules' on 62/68 (em@0 54/68); the field's clean legality of 0.999 decomposes into 0.999 on the GT-legal 95.9% plus this legalization. Pair-level remediation (field, clean) shows the gain is not silent legalization: on the 50 missing-T pairs the model *synthesizes* an unannotated T on 46 (on 23 of these a P run accompanies it, spurious on this AF excerpt and counted on the deletion audit's spurious side; of the remainder, 2 pairs lose a QRS and 2 persist locally); on the 38 skipped-QRS pairs it recovers the unannotated beat's QRS on 23 and deletes the orphan P on 15; on the non-conducted-P pairs it deletes the extra P (26/28, the remainder resolved by other local edits), the only grammar-legal option for a rhythm outside $(P\,?\,QRS\,T)^{+}$, a fidelity cost the deletion audit counts and a grammar-coverage boundary (Appendix K); the 16 split annotations all merge. Net over the 68 lead-samples: predictions carry 35 fewer P runs and 65/79 more QRS/T runs than ground truth: the pair-local fills account for 23 of the QRS and 46 of the T surplus, the rest being extra runs painted outside the census pair localities, the spurious side of the deletion audit.

Residual-illegality census of the rules variant (em@0, census of illegal transition pairs): among the 230 illegal lead-samples the rules variant still leaves, the censused illegal transitions are overwhelmingly $qrs{\to}p$ (159 pairs) and $qrs{\to}qrs$ (124 pairs): T is mandatory in the grammar, so once T is missed or dropped by the attribution filter, the following beat's P/QRS forms transitions that *reassembly cannot repair*; this corroborates the deletion audit, where rules delete the most T waves (906). Control: the bare detector's illegal transitions are dominated by T fragmentation ($t{\to}t$, 654 pairs). We also verified that 0 records produce empty output (the grammar's vacuous judgment on the empty sequence is never triggered; reclassification sensitivity is zero).

QRS-deletion edge census (em@0). QRS deletions of every variant concentrate at the evaluation-window boundaries: complexes truncated by the 4 s windows make up 5.2% of the 6,762 ground-truth QRS runs but 62–66% of the QRS deletions of the anchor (61/92), rules (64/98), and the field diffusion (101/161), a $12\times$ edge enrichment, with the deleted runs' median distance to the nearest edge at 0 samples (all runs: median 475). The interior difference is small: 0.9% (field) vs. 0.5% (rules) of non-truncated QRS missed at em@0. The mechanism is consistent with conservative decoding: a truncated complex leaves a partial bump whose shorter support is likelier to fall below the τ threshold.

Sampling rates. All signals and noise sources are resampled to the common 500 Hz of Sec. 4.1. LUDB is native 500 Hz; QTDB and AFDB (250 Hz) are upsampled $2\times$ (linear interpolation of the signal, nearest-neighbor for masks); NSTDB noise (360 Hz) is polyphase-resampled to 500 Hz before SNR scaling.

LUDB split and lead protocol. The 160/40 train/validation split of Sec. 4.1 is record-level: record identifiers are shuffled once, and the resulting lists are frozen into JSON files stored with the database directory; every later run loads the frozen lists instead of re-randomizing, so the split is identical across all variants and training seeds (the frozen lists are part of the released configuration). Under the single-lead setting, every sample is one lead of one record, all 12 standard leads are used as independent samples, and no sample is a 12-channel input: 1,920 training and 480 validation lead-

samples (the 480 quoted in Sec. 3.2). "Single lead" refers to the per-sample input, not to lead selection.

QTDB test windows. The external test of Sec. 4.1 tiles each record with non-overlapping 4 s windows starting at the record head (2,000 samples at 500 Hz; the last window is truncated at the record end), keeps a window only if its ground-truth mask contains at least one annotated wave, and emits each of the record's two signal channels as an independent single-lead sample; 103 records yield 838 retained windows and 1,676 lead-samples.

Noise construction. The *gauss* condition is additive Gaussian white noise, scaled per record to the target SNR relative to the signal's RMS; the measured conditions (bw/em/ma) use NSTDB recordings scaled to the same SNR convention.

Network and training hyperparameters. All learned variants share one backbone family and budget. The diffusion variants use a time-conditioned 1D U-Net (1.77M parameters): stem 11 filters (length 25), four down/up blocks with stride 2 and filters $23/46/92/184$ and $92/46/23/12$ (length 15, output length 9); the diffusion-noised target x_t is concatenated to the input signal channel-wise, and the time index enters every block through a sinusoidal embedding (dim 64, two-layer MLP) with FiLM modulation. The discriminative anchor is the same encoder–decoder driven with a zeroed x_t input at $t{=}0$ (4 zero channels in place of the field); the cascade control (Appendix D) splits the same budget into a coarse head plus a refinement head (1.76M total); the external baseline is a 3-filter ECG-UNet (1.84M), trained with batch-level NSTDB augmentation at our corruption recipe (bw/em/ma/gauss at SNR $\mathrm{U}[0, 12]$, 25% clean). The diffusion process is Gaussian with a cosine noise schedule ($T{=}1000$, $s{=}0.008$) and the same forward and x_0-MSE objective for every target: for the one-hot variant, Gaussian noise is added to the one-hot channels themselves and the same MSE is applied, with an identity x_0 transform at sampling and an argmax decode (analog-bits style; the focal-on-logits convention is the control of Appendix N). Training draws t independently and uniformly from $\{0, \ldots, T{-}1\}$ per sample; the $\hat{x}_0$ head is unconstrained (no clipping before the field MSE). Sampling is deterministic DDIM ($\eta{=}0$), 20 steps for test and 10 for validation, each of the K candidates an independent trajectory from fresh Gaussian noise at $t{=}T{-}1$ (no shared initialization). Capacity matching across the 3/4/6-channel targets (amplitude/one-hot/pure-phase and complex) holds at the backbone level: the same filter ladder, the channel count alone varying (a $< 0.1\%$ parameter effect). All models train with AdamW (lr $5{\times}10^{-4}$, one-cycle peak 10^{-3}; the external-baseline retrain uses $10^{-3}/2{\times}10^{-3}$), batch 32, at most 300 epochs, early stopping on validation F1 with patience 30, seed 42; three seeds $(42/43/44)$ for the main configuration, the cascade control, the one-hot diffusion (both loss conventions), the anchor (with its rules wrapper), and the discrete control of Appendix N, single seed for the remaining variants. The single-seed family members quoted in Sec. 4.4 (full 12-condition grids, unified protocol): pure-phase worst cell 0.9857 (its worst condition: clean) with F1 grid-mean 0.8613, complex worst cell 0.9922 (ma@0) with F1 0.8718. Worst-cell seed bands over those three seeds: field 0.9928–0.9952, one-hot 0.979–0.992 (focal convention: 0.928–0.968), rules 0.863–0.913, anchor 0.509–0.544. The single-draw ($K{=}1$) em@0 seed band of the field is 0.876–0.896, overlapping the rules band 0.863–0.913 (rules' seed 43 reaches 0.913, above the field's best single draw; Sec. 4.4 reads the pairing accordingly); the rules worst cell sits at em@0 for every seed, so the rules band doubles as the rules em@0 band, and the $K{=}1$ bands quoted in Sec. 4.4 are these.

Test-set label quality. On the QTDB revised-label test set, 98.2% of revised T onsets are non-regressive, and annotation agreement is markedly better than under the original labels (a cardiologist re-checked outlier beats beat by beat (Jimenez-Perez et al., 2024)).

Rule post-processing (the rules variant). The rules variant wraps the discriminative anchor's mask in a symbolic assembler with zero learned parameters; the variant with an additional template-fitting step agrees with it to within 0.1 pp on every legality/F1 reading and is not used. Steps and constants: (i) extract per-class runs; (ii) merge same-class runs separated by a gap below 120 ms; (iii) discard runs shorter than 5 ms; (iv) assemble beats anchored on QRS runs in onset order: each QRS claims the nearest unused P run whose offset precedes the QRS onset by at most 5 samples and whose onset lies within a 400 ms lookback, and the first unused T run whose onset follows the QRS offset by at most 5 samples, lies within 600 ms of the QRS onset, and ends before the next QRS onset (plus 50

samples); (v) runs attributable to no beat are discarded, and each beat's detected waves are painted back as $[\text{on}, \text{off})$ spans. The constants are fixed, physiologically motivated module defaults (max merge gap, min run length, max PR lookback, max QT search), predate the phase-field experiments, are shared by every condition and variant, and were not fitted on any evaluation split (validation or test). Steps (iv)–(v) are where deletions happen: a detected wave that fails attribution is simply gone, the "cannot create from nothing" boundary quantified in Sec. 4.3.

Legality-select tie handling. The selection score admits a degenerate tie, as an empty decode also has zero illegal transitions; no selected mask is empty in any run we audit (the beat audits of Sec. 4.5 attribute waves to every variant), and should the case arise, breaking ties toward the most decoded waves is a drop-in refinement.

Latency methodology. All latency numbers are wall-clock medians at batch 1 on one machine in one session (RTX 3080 Ti Laptop GPU), each call timed after warm-up between `torch.cuda.synchronize()` fences. Diffusion variants time the full pipeline (sampling $\rightarrow$ decoding $\rightarrow$ selection) at $K \in \{1, 2, 4, 8, 16\}$; per-candidate cost is approximately linear in K (the field diffusion: 94.8 ms at $K{=}1$, 751.9 ms at $K{=}8$). In absolute terms these process a 4 s window at 0.02$\times$ and 0.19$\times$ of real time, batched post-processing throughput, not per-beat monitoring latency. The rules end-to-end time is the anchor forward (4.4 ms median) plus the CPU assembly step measured separately over 200 records (0.8 ms median).

B Cross-condition census: one-hot vs. rules

The four-variant chain of Sec. 4.2 is measured at em@0, every variant's worst cell. Across the remaining conditions, the one-hot diffusion's advantage over rules concentrates in the heavy-noise cells: certified at em@0 ($+11.75$ $[9.21, 14.43]$), ma@0 ($+5.85$ $[4.29, 7.44]$), and em@6 ($+2.74$ $[1.53, 4.08]$), crossing zero elsewhere (clean $+0.36$ $[-0.66, 1.24]$, bw@0 $+0.60$ $[-0.72, 1.65]$, gauss@10 $+0.60$ $[-0.73, 1.95]$), unlike the field's advantage over rules, positive in every condition under every seed (Sec. 4.2). The phase-field diffusion is significantly positive in all six conditions: em@0 $+13.13$ $[10.20, 16.19]$ and clean $+1.01$ $[0.47, 1.64]$ in Sec. 4.2; ma@0 $+6.86$ $[5.24, 8.57]$, em@6 $+3.52$ $[2.08, 5.21]$, gauss@10 $+1.55$ $[0.60, 2.74]$, and bw@0 $+1.31$ $[0.71, 2.00]$ here. The representation step quoted in Sec. 4.2 (one-hot $\rightarrow$ field, $K{=}8$ select, same loss) has a positive point estimate in every condition of the grid, excluding zero in six cells (em@0 $+1.37$ $[0.48, 2.42]$ in the main text; ma@0 $+1.01$ $[0.41, 1.73]$, em@6 $+0.78$ $[0.12, 1.79]$, gauss@10 $+0.95$ $[0.06, 2.28]$, gauss@5 $+0.89$ $[0.06, 2.19]$, and ma@12 $+0.89$ $[0.12, 2.15]$ here) and touching zero in the remaining six (clean $+0.66$ $[0.00, 1.77]$, ma@6 $+0.84$ $[0.00, 2.24]$, em@12 $+0.66$ $[0.00, 1.86]$, bw@0 $+0.72$ $[-0.06, 2.07]$, bw@12 $+0.66$ $[-0.06, 1.76]$, and bw@6 $+0.60$ $[-0.06, 1.61]$); across training seeds the merged em@0 step is $+0.99$ $[0.39, 1.75]$, excluding zero (record-cluster bootstrap resampling records once per draw and averaging the three seeds' deltas on the same draw; per-seed $+1.37$ $[0.47, 2.42]$, $+0.06$ $[-0.56, 0.64]$, and $+1.55$ $[0.59, 2.80]$, seed 43 a tie whose interval crosses zero); under the focal convention the step widens to $+6.44$ $[4.55, 8.48]$ at em@0 (Appendix N). The external baseline is retrained at our early-stopping budget (patience 30); its patience-10 original collapses to 0.127 at gauss@5, a budget-driven cell that recovers to 0.519 in the retrained grid.

Table 4 gives the full grid, the data version of the heatmap of Sec. 4.2; each variant's worst cell is its column minimum, and the rules variant's heavy-noise collapse (em@0 0.863, ma@0 0.927) contrasts with the one-hot's ≥ 0.98 and the field diffusion's ≥ 0.994 in every cell. Column means reproduce the Leg. mean column of the main results table. Why a joint worst-cell interval can sit inside a single-cell interval: when a variant's two worst cells are near-tied (one-hot: em@0 0.980 and ma@0 0.986), the per-resample minimum tracks the lower envelope of two correlated statistics, and an envelope's variance is below that of either cell alone (field$-$one-hot joint $[0.49, 2.34]$ vs. the em@0 single-cell $[0.48, 2.42]$); when both minima share one cell (field and rules: em@0), the joint interval simply matches that cell's interval ($[10.15, 16.22]$ vs. $[10.20, 16.19]$). All joint intervals are record-cluster bootstraps, resampling records once per draw and recomputing both grids' minima on the same draw.

Table 4: Full legality grid (12 conditions × five variants; external QTDB test); the worst-cell row is each variant's column minimum. "+ field" is the phase-field diffusion of Sec. 3.2, the amplitude implementation of the family ("the field"); the one-hot column is the x_0-MSE convention of the main comparison (focal convention: Appendix N).

Condition	Ext. U-Net	Anchor (bare)	+ Rules	+ one-hot	+ field (ours)
clean	0.773	0.783	0.989	0.993	**0.999**
gauss@5	0.519	0.742	0.982	0.991	**0.999**
gauss@10	0.704	0.794	0.983	0.989	**0.999**
bw@0	0.740	0.747	0.986	0.992	**0.999**
bw@6	0.757	0.764	0.989	0.993	**0.999**
bw@12	0.761	0.779	0.991	0.993	**0.999**
em@0	0.488	0.544	0.863	0.980	**0.994**
em@6	0.650	0.680	0.964	0.991	**0.999**
em@12	0.709	0.708	0.988	0.993	**1.000**
ma@0	0.502	0.579	0.927	0.986	**0.996**
ma@6	0.682	0.718	0.980	0.992	**1.000**
ma@12	0.730	0.764	0.985	0.991	**1.000**
worst cell	0.488	0.544	0.863	0.980	**0.994**

C FACTOR CERTIFICATION: THE TARGET SIDE CARRIES COLLAPSE AND AVERAGABILITY

A 2×2 anti-diagonal experiment decouples the conditioning side from the target side. Amplitude-conditioned $\rightarrow$ one-hot-target still collapses (mean worst cell 0.18), while one-hot-conditioned $\rightarrow$ amplitude-field-target does not (mean 0.83). The target-side gain at em@0 is $+1.01$ $[0.12, 2.07]$, excluding zero, and the conditioning side crosses zero. Two readings must not be conflated: this $+1.01$ is measured on the deterministic coarse-head path of the cascade family, whereas the diffusion path's target step of Sec. 4.2 is $+1.37$ at $K{=}8$ selection ($+6.44$ under the focal convention; Appendix N), same direction, different operating point. Collapse and averagability are target-side properties relative to the network; on the decode axis, the collapse is specific to the decision-boundary readout (Appendix P). The detection-cascade variant of Appendix D reads the same 2×2 grid along the deployment axis and certifies the same conclusion from the conditioning side: external anchoring adds $+4.47$ $[2.86, 6.19]$ to the one-hot path but nothing significant to the field (at an F1 cost).

D CASCADE DEPLOYMENT VARIANT: FULL COMPARISON

The cascade variant places the diffusion second in a detection cascade: a coarse prediction head, trained for wave localization on the same corrupted inputs, runs one deterministic forward whose output conditions a refinement diffusion head that samples fields. The single-network main configuration retrains the same backbone end to end with the diffusion objective, without the coarse head. This appendix carries the full-dimension comparison summarized in Sec. 4.4; the main-text tables omit the cascade rows.

Legality and F1. Under the amplitude-field target, adding the cascade contributes nothing significant to legality in any of four conditions (em@0 $+0.12$ $[-0.30, 0.57]$) and costs F1 -0.75 to -0.87 (excluding zero; em@0 -0.75 $[-1.32, -0.22]$); on the one-hot path, in contrast, the cascade adds $+4.47$ $[2.86, 6.19]$: the cascade mainly rescues one-hot (by providing wave-localization anchors), while the amplitude field barely needs it. Grid readings for the cascade variant: F1 0.874, legality mean 0.997, worst cell 0.995.

Sampling and aggregation robustness. The cascade's exclusive value: $K{=}1$ 0.939 (vs. 0.888 for the single network), mean@2 0.890 (vs. 0.826), mean@16 0.932 (vs. 0.841).

Beat-level audit and latency. Beat audit at em@0: full 0.756, full_cyc 0.639, qt_usable 0.946, frag 0.023, frag_cyc 0.154; the single network has the higher paired full in 4 of 5 conditions (point estimates, e.g., clean $+0.99$; the record-cluster CIs cross zero throughout, e.g. clean $[-0.53, 2.72]$),

tying the fifth. Latency (batch 1, medians, same machine and session): 102 ms at $K{=}1$ and 773 ms at $K{=}8$; the single network is slightly faster at small K (95 vs. 102 ms), and the two are comparable at large K (752/773 ms).

Decision. We adopt the single network as the main configuration: simpler structure (one forward path, no coarse head), higher F1, on-par legality, lower small-K latency. When extreme single-draw quality is required, the cascade is an optional enhancement.

E Robustness scans: decode threshold and tolerance

τ scan (decode threshold). Lowering the decode threshold τ from 0.35 to 0.05 recovers only 0.9 pp of P recall, with F1 and legality nearly unchanged: the clean-condition misses located in Sec. 4.5 are "field-did-not-paint" omissions rather than a threshold artifact, and the one-hot shares them (clean P recall 0.822 vs. the field's 0.818, its argmax decode having no threshold to scan).

Tolerance scan. F1 at 75 ms tolerance under em@0: 0.806 vs. rules' 0.792; at 150 ms: 0.861 vs. 0.846. The ordering is preserved at both tolerances.

Channel-held-out noise realization. The corruption block and the test grid both draw their measured noises from NSTDB, so training and test noise are never fully independent. The audit here keeps the trained model fixed (corruption on channel 0) and switches the grid to the channel-1 recordings, a noise realization never applied during training: grid-mean legality changes by -0.05 pp (0.9987 $\rightarrow$ 0.9982; worst cell 0.994 $\rightarrow$ 0.993), so the gain does not rest on test-noise identity with the corruption source; artifact families beyond NSTDB's three plus Gaussian remain outside this audit.

F Real-AF external test (AFDB)

Legality is a GT-free metric, evaluable directly on unlabeled real signals; this appendix is the deployment-grade demonstration summarized in Sec. 4.2. MIT-BIH AFDB (Moody & Mark, 1983; Goldberger et al., 2000) (23 records of $\sim$10 h) ships rhythm-episode annotations; we cut 10 s windows from authority-annotated AF episodes and non-AF episodes alike (at most 6 windows per episode, per-lead samples: 2,796 AF / 3,312 non-AF windows) and compare three variants: bare anchor, rules, ours.

Legality (AF episodes, clean $\rightarrow$ em@6). Bare anchor 0.16 $\rightarrow$ 0.12, rules 0.81 $\rightarrow$ 0.70, ours 0.88 $\rightarrow$ 0.86; the gain over rules has record-cluster CIs excluding zero (clean $+6.8$ $[3.1, 12.9]$; em@6 $+15.3$ $[8.1, 25.3]$; 20 records), and the gap widens under noise. On this P-less rhythm from a database disjoint from training, the legality gain carries an annotation-anchored QRS audit (below) and empirically justifies the grammar's P-optional design of Sec. 3.1.

P-wave audit (anti-gaming). Legality does not punish hallucinated P ($P{\rightarrow}QRS$ is legal): the predicted P/QRS ratio on AF segments is 0.21 / 0.20 / 0.17 for the three variants, i.e., *true omission* rather than per-beat hallucination; the high legality is not bought by hallucination. On non-AF (sinus) segments the ratios are 0.78 / 0.73 / 0.63: when P is present, it is drawn; the representation conditions P on presence.

R-wave GT-proxy audit. AFDB ships beat-level `.qrs` annotations, so the GT-free legality above can be anchored externally: each predicted QRS run is matched against the annotated R peaks of the same window ($\pm$150 ms). All three variants place QRS on real beats (median $|\Delta t|$ 10–12 ms in every condition), and the field diffusion has the *fewest* unsupported QRS runs throughout (AF episodes, clean $\rightarrow$ em@6: 0.73% $\rightarrow$ 0.78%, vs. rules 1.00% $\rightarrow$ 1.20% and the bare anchor 1.13% $\rightarrow$ 1.40%): the high legality is anchored to annotation, not hallucinated regularity. The honest cost is coverage: the field recovers 97.2–97.9% of annotated R peaks on AF episodes (rules 98.9%), the same paint conservatism the τ scan locates in Appendix E; hallucinated T waves remain unauditable at this proxy level, as no T annotations exist in AFDB.

Built-in control. Same-record non-AF episodes serve as a control (rules 0.93/0.89, ours 0.955/0.946), confirming that the collapse and recovery on AF episodes is a *rhythm effect*, not domain shift. All CIs in this appendix resample records (the 10 s windows of a record form one cluster; 20 AF / 21 non-AF records carry usable windows), matching the record-level protocol of the main text; an earlier revision reported window-level intervals, which were anti-conservative under within-record correlation.

G BEAT-LEVEL AUDIT: VERDICTS, DUAL TRACKS, RESIDUALS

Verdicts. A beat is a ground-truth QRS run; its expected waves (P and T) are those present in ground truth for that beat. Each predicted run is attributed to the ground-truth run with maximal same-class overlap; the beat's *anchor* is its attributed QRS run. The six verdicts, in order of severity:

- *miss*: no attributed QRS within 150 ms of the ground-truth QRS onset;
- *frag*: the anchor is present, but some class has more than one attributed run, or an attributed P starts at or after the anchor's onset (or T at or before it), or an unattributable spurious run overlaps the beat's span;
- *partial_qt*: the expected P is missing while QRS (and T, if expected) are fine, so the QT interval remains measurable;
- *partial_pr*: the expected T is missing, so the PR interval remains measurable;
- *partial_none*: both expected P and T are missing;
- *full*: every expected wave paired exactly once, correctly ordered, with no spurious run inside the span.

qt_usable = full + partial_qt; pr_usable = full + partial_pr. Verdicts inspect structure only, not boundary accuracy (boundaries belong to the precision layer).

Single-draw rows, the external baseline, pool consistency, and per-record CIs. Single-draw ($K{=}1$) rows: one-hot 0.701/0.595/0.891/0.054/0.180, field 0.757/0.638/0.928/0.031/0.167; external baseline: 0.675/0.544/0.863/0.116/0.272. With per-batch seeding fixed, the $K{=}1$ masks are the first candidates of the pool the $K{=}8$ rows select from, so the two operating points are strictly pool-comparable. Record-cluster paired CIs (ours − rules): em@0 $+2.63$ $[0.97, 4.38]$ and em@6 $+2.60$ $[0.87, 4.47]$ exclude zero; gauss@10 $+1.25$ $[-0.69, 3.45]$, ma@0 $+0.11$ $[-1.62, 1.92]$, and clean -1.00 $[-3.08, 1.18]$ cross zero (the clean reversal of Sec. 4.5 is thus a point estimate, significant only under the narrower lead-sample CI). Per-condition full at $K{=}8$ (one-hot vs. field): clean 0.829 vs. 0.827, gauss@10 0.791 vs. 0.803, em@6 0.778 vs. 0.800, em@0 0.725 vs. 0.763, ma@0 0.776 vs. 0.792: the one-hot trails in every noisy condition with a clean tie, and its $K{=}1$ full barely clears the bare anchor (0.701 vs. 0.694), the readings quoted in the lead of Sec. 4.5; under the focal convention the light-noise comparison flips in one-hot's favor (4 of 5 conditions, field−one-hot clean -1.83 $[-3.67, -0.16]$, gauss@10 -2.77 $[-4.90, -0.91]$, significant under both; Appendix N). Per-record absolutes at em@0: rules 0.742, ours 0.768. On the cycle track the same per-record pairing (ours − rules) is -2.95 $[-4.82, -0.99]$ at em@0 and -8.32 $[-10.37, -6.17]$ under clean (the full_cyc values quoted in Sec. 4.5).

Dual tracks. The same verdicts are counted over two spans. The *measurement track* takes the union of the beat's ground-truth runs, answering "can intervals be measured over this signal segment." The *cycle track* takes the span from the previous beat's T offset to this beat's T offset, answering "is this cardiac cycle clean": TP-segment ghost waves that overlap no ground-truth measurement span still fall inside cycle spans and count against the beat as *frag*; this is the inter-beat tax discussed in Sec. 4.5. There are 6,762 ground-truth beats per condition. Main-text audit tables are pooled over beats; paired CIs use the per-record basis, and the two differ by at most 1 pp (rules at em@0: pooled full 0.738 vs. per-record 0.742).

Residual census of the field diffusion (em@0, the main configuration): among the 10 illegal lead-samples (0.6%), the illegal-transition census is $qrs{\to}qrs$ (5), $qrs{\to}p$ (2), $t{\to}t$ (2), $p{\to}t$ (1): T-miss-type residuals isomorphic to the rules variant's census of Appendix A, plus fragmentation

($t \to t$) concentrated on short segments; the 10 lead-samples carry 1–7 ground-truth beats (median 3.5).

H INTERVAL ERROR TABLES

Table 5: QTDB intervals: median absolute error (ms) / paired-beat count on ground-truth-paired beats (QRS anchor within 200 ms); ours vs. the rules variant.

	PR		QT		QRS	
Condition	Ours	Rules	Ours	Rules	Ours	Rules
clean	16 / 4,584	**14 / 4,620**	20	**18**	**12**	**12**
gauss@10	20 / 4,481	**16 / 4,455**	22	**20**	**12**	**12**
em@6	20 / 4,536	**18 / 4,420**	24	**22**	**12**	**12**
em@0	**26 / 4,330**	**26 / 4,091**	28	**24**	**14**	**14**
ma@0	20 / 4,441	**18 / 4,393**	24	**20**	14	**12**

Table 5 reports the per-condition QTDB medians and paired-beat counts referenced in Sec. 4.5. The in-distribution reference for that cross-domain comparison is Table 6: median absolute interval errors on the LUDB validation split. The external U-Net baseline edges the field diffusion by 2–4 ms throughout (PR 6/10/14/12 vs. 8/14/18/14); both roughly double on QTDB, the cross-domain effect of Sec. 4.5.

Table 6: LUDB validation intervals (median error, ms; conditions clean/em@6/em@0/ma@0).

Variant	PR	QT
External U-Net baseline	**6 / 10 / 14 / 12**	**8 / 10 / 14 / 10**
Amplitude-field diffusion (ours)	8 / 14 / 18 / 14	10 / 14 / 16 / 14

Paired-beat counts (coverage). Under noise the field pairs the most beats (em@0 PR: 4,330 vs. rules 4,091 and bare detection 3,939); under clean, rules pair slightly more (4,620 vs. 4,584).

I CORRUPTION IS NECESSARY

The in-model corruption block is not optional. Retraining the focal-convention one-hot control (Appendix N) without it (same 1.77M backbone, same optimizer, schedule, and early stopping, clean inputs only) collapses the worst cell from 0.928 to 0.334 (legality mean 0.966 $\to$ 0.809, F1 0.877 $\to$ 0.846): a prior trained only on "clean signal $\to$ field" has nothing to restore when the test input is noisy. The same holds for the field target at the cascade level (worst cell 0.615 without corruption vs. 0.995 with it), so the corruption dependence is a property of the generative head, not of the target representation.

J GENERATIVE VERSUS DETERMINISTIC HEADS

This appendix separates the contribution of the generative head from that of the target representation, under the shared capacity, corruption, and training budget. Table 7 collects the variants: the discriminative anchor (cross-entropy classification), two deterministic-head cascade controls (cross-entropy refinement over one-hot or amplitude conditioning), a *deterministic regression* variant that trains the single-network backbone to regress the amplitude field directly with MSE (no diffusion process, a single forward at test), and the diffusion variants.

Three readings. **First, under a classification head the representation effect is absent**: swapping the cascade's conditioning from one-hot to amplitude moves em@0 by -3.5 pp (0.568 vs. 0.603): cross-entropy refinement has no sampling manifold and nothing to average, so the field buys nothing there. **Second, deterministic regression onto the field keeps a third of the diffusion gain**: worst cell $+13.8$ pp over the anchor (0.682 vs. 0.544), clean $+13.5$ pp (0.918 vs. 0.783), and the best F1 of

Table 7: Head × target attribution (external QTDB test, 12-condition grid; deterministic variants are single forwards, diffusion variants at $K=8$ select).

Variant	Head / target	Leg. worst	Leg. mean	F1@150 ms	em@0
Anchor	CE, one-hot	0.544	0.717	0.865	0.544
Cascade control	CE refine, one-hot cond.	0.597	0.735	0.864	0.603
Cascade control	CE refine, amplitude cond.	0.568	0.748	0.871	0.568
Regression	MSE, amplitude field	0.682	0.868	**0.885**	0.682
Diffusion	sampling, one-hot	<u>0.980</u>	<u>0.990</u>	0.875	<u>0.980</u>
Diffusion (ours)	sampling, amplitude field	**0.994**	**0.999**	<u>0.880</u>	**0.994**

all variants (0.885); the representation carries boundary precision and a legality floor even without sampling. **Third, generation supplies the ceiling**: at the same backbone, target, and protocol, diffusion sampling adds $+31.2$ pp over regression (0.994 vs. 0.682): under heavy noise the posterior over fields is multimodal (a low-amplitude wave is either present or absent), and the conditional mean smears instead of committing, so missed waves stay missed; this is the temporal-structured counterpart of the classical inadequacy of mode and mean decoding, for which sampling and minimum-Bayes-risk decoding are the standard remedies (Kumar & Byrne, 2004; Eikema & Aziz, 2020). The deterministic regressor is thus the exact dual of the one-hot diffusion: best F1 with weak structure (0.885/0.682) versus salvageable structure with fragile single draws (0.875 at $K=8$/0.794 at $K=1$). Paired-bootstrap CIs (record-cluster; all legality intervals exclude zero): the regression-vs-anchor step is $+13.78$ $[10.47, 17.08]$ at em@0, $+13.48$ $[9.78, 17.35]$ under clean, $+18.32$ $[15.27, 21.46]$ at em@6, $+18.44$ $[15.27, 21.70]$ at ma@0; the diffusion-vs-regression step is $+31.21$ $[27.05, 35.35]$ at em@0, $+8.11$ $[4.65, 12.01]$ under clean, $+13.54$ $[9.77, 17.72]$ at em@6, $+23.27$ $[19.53, 27.20]$ at ma@0. On F1 the two steps act differently: regression improves it everywhere ($+1.4$ to $+2.7$ pp, excluding zero), while sampling is F1-neutral-to-slightly-negative (-1.03 $[-1.68, -0.43]$ under clean, $+0.31$ $[-0.39, 1.00]$ at em@0): the regressor's F1 edge is a clean-condition phenomenon, and legality is what sampling buys. The head swap itself at the anchor level (CE → diffusion, same non-cascade backbone) moves F1 by $+2.23$ $[1.46, 3.05]$ per-record at em@0, the generative-head step quoted in Sec. 4.2. On the focal basis the same representation step is $+0.80$ $[0.12, 1.47]$ per-record at em@0; the pooled point estimate of $+0.5$ is the main-table grid-mean difference (0.880 vs. 0.875).

K LIMITATIONS, IN ORDER OF IMPACT

The two most consequential are summarized in the conclusion; all nine are listed here with their evidence and mitigations.

1. **Inter-beat looseness**: the cycle track is below rules in all conditions (1.99 TP ghost waves per record); inter-beat consistency of the diffusion head is the primary improvement direction. The stacked assembler of Appendix O is a measured stopgap ($+6.6$ pp cycle-track full, CI excluding zero, at 0.8 ms CPU; T deletions $\times 1.8$, F1 -1 pp); reproducing it inside the head (cycle-aware sampling) is the concrete target.
2. **Clean conservatism**: clean full -1.00 (rules higher; the record-cluster CI crosses zero, so the gap is a point estimate); the τ scan locates it as "field-did-not-paint" misses, i.e. upstream conservatism rather than a decoding-threshold problem.
3. **QRS deletions** $+63$: rules are better (98 vs. 161); two-thirds of both counts are window-edge truncations shared by all variants ($12\times$ edge enrichment, Appendix A), the interior gap at em@0 is 0.9% vs. 0.5% of non-truncated QRS, and conservative decoding over-shrinks the partial bumps of truncated complexes.
4. **The representation gain depends on the generative head**: not reproduced under a discriminative CE head (-3.5 pp at em@0, Appendix J); a deterministic regressor onto the field keeps a third of the diffusion gain with the best F1 (0.885) but saturates at 0.682, below rules; the legality claim is restricted to generative structured prediction.
5. **Rhythm coverage**: AF-episode legality 0.88/0.86 (clean/em@6; Appendix F) does not reach the 0.99 tier of the QTDB grid, a data-side rather than structure-side gap; the fix (training-data composition) is beyond this paper's scope.

6. **Grammar rhythm coverage**: the transition grammar cannot express non-conducted P sequences (second-degree block), ventricular rhythms without a resolvable T, or paced rhythms. On the ground-truth-illegal census of Appendix A, the legality gain over ground truth on these lead-samples comes partly from deleting the extra P (26/28 pairs): the metric rewards the deletion on exactly the conduction-disorder population whose PR measurement matters most. Extending the grammar (optional-T or P-run admission under block) or reporting legality stratified by rhythm is future work.

7. **External-test surface**: QTDB with revised labels is the only external test with wave-boundary ground truth; AFDB adds GT-free legality with an R-wave annotation proxy audit (Appendix F; no T/boundary ground truth); both single-lead 500 Hz.

8. **Single-task evidence base**: all evidence comes from one task (ECG delineation), one grammar ($(P\,?\,QRS\,T)^+$), one backbone family, and one signal modality; the representation-level lesson is conjectured to transfer to other structured-prediction tasks under grammar or topology constraints (Sec. 5). Beyond this surface, a single-seed heart-sound probe replicates the single-draw and mean@2 stability conclusions, with a regime boundary on large-K averaging (Appendix Q); broader multi-domain, multi-seed verification remains future work.

9. **Residual illegality**: em@0 legality 0.994 does not reach 1. The residual 10 illegal lead-samples (0.6%) are T-miss-type residuals isomorphic to the rules variant's, plus fragmentation residuals on short segments (1–7 ground-truth beats); the census is in Appendix G.

L MECHANISTIC ILLUSTRATIONS

This appendix collects three illustrations referenced from the main text. Target encoding (Fig. 4): one beat under the two targets, one-hot flat blocks versus the amplitude field's graded channels. Averaging (Fig. 5a): two misaligned one-hot samples average point-wise to an intermediate plateau, and per-point argmax decoding flips repeatedly near the boundary, producing fragments interleaved with other classes, an illegal sequence: this side is exact, as argmax does not commute with point-wise averaging. Two misaligned $\sin(\pi\phi)$ bumps average to a field of lower amplitude; within a single run and misalignment below $\approx 2\phi_\tau d$ (the τ-margin at both ends), the τ decode stays legal and its support interpolates the two candidates; larger misalignment can split the support into two same-class runs or, for weakly painted waves, drop it below τ entirely. Field-side robustness under averaging is therefore an empirical regularity (mean@2 legality 0.826, below the single draw's 0.888), not a closure property. Closed-form decoding (Fig. 5b): from the thresholded support $[s_\ell, s_r]$ the run duration d and the true boundaries on/off follow analytically from the shape of $\sin(\pi\varphi)$.

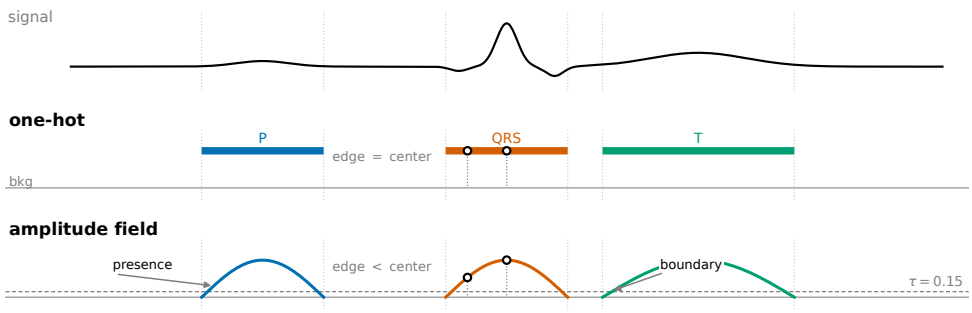


Figure 4: Run encoding of one beat under the two targets (top: the signal). Middle: the one-hot target, where each wave class is a flat block on a shared row (color = class); the two markers (near-edge and center of QRS) receive the same value. Bottom: the amplitude field, where each class occupies a channel (color = class) whose value traces $\sin(\pi\phi)$; the same two samples receive distinct, graded values. Presence is a non-zero channel value; boundary is the support end above τ; order is carried by the channel (color).

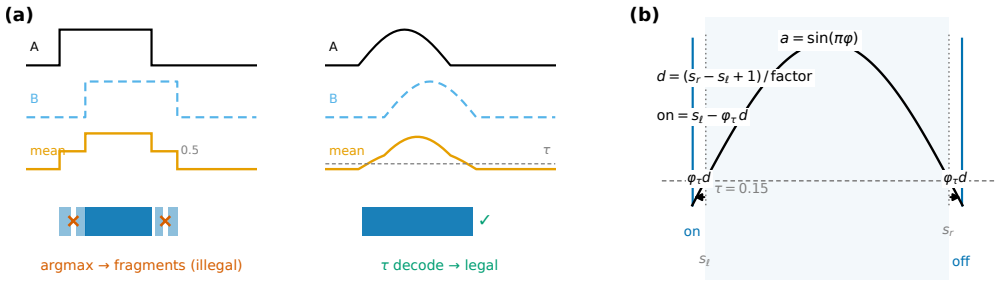


Figure 5: Mechanistic illustrations. (a) Point-wise averaging: breaks one-hot exactly (argmax $\rightarrow$ fragments, illegal); the field's τ decode stays legal at the misalignment shown, its support interpolating the two candidates. (b) Closed-form decoding: the support $[s_\ell, s_r]$ above τ determines duration d and boundaries on/off by analytic extrapolation.

M Qualitative comparison

Fig. 6 shows one external-test record under electrode motion at 0 dB (em@0): the bare anchor's mask fragments, rule post-processing deletes waves, a single one-hot draw fragments, and the phase-field diffusion stays intact; the ground truth is overlaid as shaded spans on the noisy signal in the top panel.

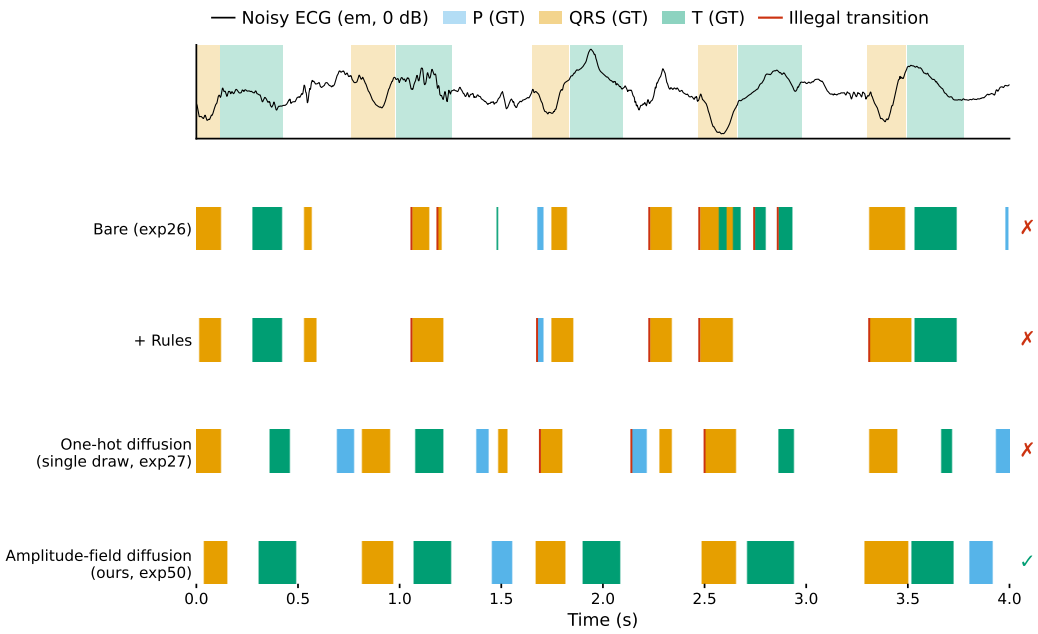


Figure 6: Qualitative comparison (em@0, one record): mask timelines of four variants on the same signal; ground truth is overlaid as shaded spans on the noisy signal in the top panel. Read as a tracing: the anchor and one-hot masks break the $P \rightarrow QRS \rightarrow T$ sequence, leaving PR and QT unreadable on the affected beats; rules' deleted waves take their intervals with them; the phase-field mask keeps P, QRS, and T in sequence on every beat, so both intervals stay measurable.

N Focal-on-x_0 and discrete diffusion do not rescue one-hot

The main comparison of Sec. 4.2 holds the training loss fixed: both targets train the same Gaussian x_0-MSE. Two counter-readings remain, each tested here by a control that swaps exactly one axis. The generative-family reading: discrete diffusion exists precisely because categorical targets resist

Gaussian corruption (Austin et al., 2021; Lou et al., 2024), so perhaps the observed fragility reflects a mismatched sampler rather than the target. The convention reading: the focal-on-x_0 convention (the DiffAct family's default) differs from the main one-hot in both loss and readout, so perhaps the main comparison overstates what that convention achieves. Both controls hold everything else fixed.

Setup. An absorbing-state discrete diffusion (Austin et al., 2021) over the same 4-class one-hot target, on the same DenoiseUNet backbone (parameter ratio 1.0002 with the Gaussian variant), with the same corruption block, optimizer, one-cycle schedule, early stopping on validation F1, training seeds 42/43/44, and 20 denoise steps at inference; training minimizes focal loss on the predicted x_0 distribution (the readout also used at inference, configuration `exp57` of the released code). The forward process absorbs each position into an extra [MASK] state with probability $1 - \bar{\alpha}_t$ on the cosine table shared with the continuous variants; the reverse is ancestral posterior sampling: a masked position stays masked with probability $(1 - \bar{\alpha}_{t-1})/(1 - \bar{\alpha}_t)$ and is otherwise drawn from the predicted x_0 distribution, revealed positions frozen. Candidates are read out per trajectory as the x_0 argmax (the D3PM-style readout), the exact analogue of the Gaussian variant's per-trajectory logits: per-candidate argmax, mean-then-argmax aggregation, unified seeding protocol, 16-candidate pool. Scope: the absorbing kernel is the one D3PM transfer kernel tested; the uniform kernel and score-entropy training (Lou et al., 2024) are untested, and the 20-step budget matches the Gaussian variant's competent regime rather than the full-length (T-step) sampling under which absorbing kernels are usually evaluated.

Table 8: The one-hot target under the main x_0-MSE convention and under two controls (focal-on-x_0; absorbing-state discrete diffusion); same backbone, capacity, corruption, and budget; operating point $K{=}8$ select; F1 at 150 ms, grid mean.

Variant	em@0			Leg. worst	Leg. mean	F1
	$K{=}1$	$K{=}8$	mean@2			
Gaussian one-hot (x_0-MSE, main)	**0.794**	**0.980**	0.000	**0.980**	**0.990**	<u>0.875</u>
Gaussian one-hot, focal (DiffAct)	<u>0.349</u>	<u>0.930</u>	**0.009**	<u>0.928</u>	<u>0.966</u>	**0.877**
Discrete one-hot (D3PM, control)	0.171	0.673	<u>0.006</u>	0.673	0.878	0.857

Readings (Table 8; record-cluster CIs as in Sec. 4.1). **First, single-draw robustness is not rescued but worsened**: 0.171 at em@0, 69.2 pp below rules ($[-72.15, -66.14]$) and 17.8 pp below even the focal-convention Gaussian one-hot ($[-20.76, -14.80]$). The mechanism is the sampler's commitment ordering: absorbing diffusion reveals positions early and freezes them, so a misaligned early reveal conditions every later step, whereas DDIM re-estimates the whole field at each step and reads out one globally consistent decode. **Second, selection does not reach rules**: $K{=}8$ 0.673 vs. 0.863 (-18.97 $[-21.96, -15.91]$), and the pool itself is poorer: at $K{=}16$ a legal candidate exists for only 82.9% of lead-samples (focal Gaussian one-hot: 98.2%). The $K{=}2$ pool rates quoted in Sec. 4.4 (em@0: field 95.8%, x_0-MSE one-hot 90.5%, focal one-hot 56.9%) come from the same unified 16-candidate scan: the focal convention's single-draw fragility empties the early pool, and only a large K recovers it. **Third, the mean collapse is identical**: mean@2 0.006 vs. 0.009 (-0.30 $[-0.77, 0.18]$, crossing zero), as the decode geometry predicts: misaligned one-hot samples average to the same intermediate plateau whichever sampler produced them. **Fourth, the tier is not a sampler-budget artifact, within the tested budget**: doubling the denoise steps to 50 moves $K{=}8$ only from 0.673 to 0.705, and the pure-realization readout (the sampled token sequence itself) never yields a legal record at em@0 (0.000 at every K): categorical realization exposes posterior entropy as per-position boundary noise, fragmenting even clean signals by an order of magnitude above the ground-truth run count. The residual rise at 50 steps is compatible with slow ancestral convergence, so full-length sampling could narrow the gap further; we did not run it, and read the exclusion at the 20/50-step tier as budget-conditional. **Fifth, the deletion audit locates the failure**: the control deletes the fewest ground-truth P (1,264) and T (466) waves of all variants yet produces the most spurious runs of all (P 1,823, QRS 1,909, T 2,358): the discrete family supplies presence even more liberally than the Gaussian, and the one-hot decode converts that surplus into illegal fragmentation. Generation supplies waves; the representation supplies the sequence.

Focal-convention control (one-hot + focal-on-x_0, the DiffAct default). The remaining axis is the readout convention itself: the DiffAct family trains focal loss on the predicted x_0 logits and reads out

a single deterministic trajectory. This control retrains the Gaussian one-hot under that convention (configuration `exp27`); the main one-hot of Table 1 is its x_0-MSE counterpart (analog-bits style: MSE on the one-hot channels, identity x_0 transform at sampling, argmax decode; configuration `exp58`, seed 42). The convention costs single-draw robustness most: at em@0 the single draw falls from 0.794 to 0.349 (51.4 pp below rules) and the worst cell from 0.980 to 0.930 (Table 8); F1 is unchanged (0.877 vs. 0.875): the swap moves structure without moving precision. The convention step (focal $\rightarrow$ MSE, $K{=}8$) is certified in every condition (em@0 $+5.07$ $[3.50, 6.81]$; ma@0 $+5.73$ $[4.06, 7.49]$; clean $+1.49$ $[0.69, 2.39]$). What the focal convention does retain are two light-noise edges, both conventional: its beat-level full exceeds the field's in 4 of 5 conditions (field$-$focal clean -1.83 $[-3.67, -0.16]$, gauss@10 -2.77 $[-4.90, -0.91]$, significant under both; Appendix G), and its P recall at em@0 is the highest of all variants (0.790) at the cost of the most spurious P and T runs among repaired variants (T 1,666 vs. the main one-hot's 1,283 and the field's 1,215); the loss match erases both (Secs. 4.3, 4.5). Under focal the representation step at $K{=}8$ widens to $+6.44$ $[4.55, 8.48]$ at em@0 (positive in all audited conditions, Appendix B): part of the focal-basis gap is the loss, not the target. The loss axis is probed where it binds: the one-hot side. The field is a regression target whose x_0-MSE convention is its natural pairing (focal loss is defined for classification logits), so no field-side convention swap is tested.

Seed robustness. Across the three training seeds the D3PM control's worst cell spans 0.201–0.673 ($K{=}8$) against rules' 0.863–0.913 and the main one-hot's 0.979–0.992: below both bands at every seed, with an instability the reference variants do not show. Its em@0 single draw spans 0.029–0.171 and mean@2 collapses at every seed (0.000–0.006), while grid-mean F1 stays in the 0.84–0.86 tier: precision survives; legality is what collapses. Seed 42, shown in Table 8 and Fig. 3, is the D3PM control's *strongest* seed, so the quoted operating points are its most favorable draws and the exclusion is conservative. The main one-hot (x_0-MSE) is seed-stable: across the same three seeds its em@0 single draw spans 0.762–0.864 (at best a rounding-level tie with the rules floor, 0.863–0.913) and its worst cell 0.979–0.992, both bands disjoint from the field's (single draw 0.876–0.896, worst cell 0.9928–0.9952; the paired worst-cell step against the field is $+1.37$ $[0.49, 2.34]$, joint record-cluster CI; under the focal convention the joint worst-cell step against the field is $+6.56$ $[4.98, 8.75]$, the interval quoted in Sec. 4.2), and mean@2 is 0.000 at every seed and condition. The focal control spans 0.928–0.968 at the worst cell, and its one-hot-vs-rules ordering under clean and light noise is seed-sensitive (seed 42: rules significantly better under clean and bw@0; seeds 43/44: the signs reverse by $+1.3$ to $+2.4$ pp); the x_0-MSE census of Appendix B carries no such sensitivity.

Attribution. With Appendix C (network factorization), Appendix D (deployment conditioning), Appendix J (deterministic heads), the D3PM control (the nearest discrete generative family), the focal-convention control above, and the decoder-symmetry control (Appendix P), every axis of the pipeline (sampler family, training loss, readout convention, decode geometry) has been swapped except the target representation itself: the collapse follows the target through all of them, and the single-draw deficit shrinks under the friendliest loss yet never closes (0.794 vs. rules' 0.863). Within the tested scope (an absorbing-state kernel at 20/50-step budgets; a focal-on-x_0 convention swap), the readings "the right discrete generative family would rescue one-hot" and "the right loss would" find no support; untested alternatives remain, notably other transfer kernels (uniform, score-entropy) and discrete vocabularies that do carry within-wave position (class $\times$ phase-bin tokens, the analog-bits direction (Chen et al., 2023)), which would probe the position diagnosis rather than the continuity prescription directly.

O RULES STACKED ON THE GENERATIVE VARIANTS: REPAIR ABSORBED, CURATION RETAINED

Sec. 4.2 measures the rules variant on the anchor's output and the diffusion variants bare; this appendix completes the matrix by running the same assembler of Appendix A (zero learned parameters, constants fixed before the phase-field experiments and never refitted) on the decoded masks of the diffusion variants, both one-hot conventions included. All arms reuse the main grid's seeds, noise realizations, and protocol; the bare arms reproduce the archived grids bit-exactly (field arms: three seeds, all 12 conditions; one-hot arms: seed 42, the audited cells).

Table 9: Assembler lift at em@0 by input (seed 42; legality and F1, bare vs. stacked).

Input to the assembler	Leg. bare	Leg. +rules	Lift (pp)	F1 bare	F1 +rules
Discriminative anchor	0.544	0.863	+31.9	0.827	0.846
One-hot, focal conv. (K=8 select)	0.930	0.989	+6.0	0.849	0.848
One-hot, x_0-MSE conv. (K=8 select)	0.980	0.984	+0.4	0.849	0.845
Field diffusion, K=1 single draw	0.888	0.937	+4.8	0.858	0.852
Field diffusion (K=8 select, ours)	0.994	0.996	+0.2	0.861	0.854

Reading 1: repair is absorbed. On the field's main operating point the assembler is a bounded null on the primary metric (Table 9): the em@0 delta (both variants' worst cell at seed 42) is $+0.18$ $[-0.06, 0.52]$ pp (record-cluster CI crossing zero), the clean delta is exactly zero (not one record changes), the seed-42 per-condition deltas lie within $[-0.06, +0.18]$ pp (largest across seeds $+0.48$, seed 43 at em@0), and the worst-cell seed band shifts from 0.9928–0.9952 to 0.9958–0.9970. The residual census localizes the repair: exactly the three fragmentation-type residuals of Appendix G ($t\to t$ $\times 2$, $p\to t$ $\times 1$) disappear, while the seven T-miss residuals ($qrs\to qrs$ $\times 5$, $qrs\to p$ $\times 2$) survive, the assembler's own failure signature (Appendix A). The repair duty has moved into the representation; rules have nothing left to repair on this output.

Reading 2: what survives is curation, and curation is deletion. The one axis where the stack helps is the cycle track: full_cyc $+6.58$ $[5.51, 7.63]$ at em@0 and $+7.08$ $[5.95, 8.23]$ under clean (both record-cluster CIs excluding zero), uniform across all 12 conditions ($+5.1$ to $+5.9$ pp pooled) and across seeds; TP ghost waves drop $1.99 \to 1.59$ per record (-20%; the rules arm's own tier is 1.73). The measurement track pays nothing: full is flat in every condition (em@0 -0.06 $[-0.23, 0.07]$), and of the $+403$ additional T-run deletions not one creates a new partial beat (5,167 $\to$ 5,166 full beats, zero new partial-pr verdicts): the deletions land on already-broken beats and ghost runs. The cost is precision and deletions: F1 -0.7 to -1.2 pp in every condition, T deletions double (495 $\to$ 898 at em@0) at an input-independent appetite (906 on the anchor, 981 on the focal one-hot, 974 on the x_0-MSE one-hot, 898 on the field), and P recall 0.754 $\to$ 0.738. Deleting cures extras; generating cures misses; the field has already internalized everything deletion could cure.

Reading 3: the one-hot columns and the cheap path. On the matched (x_0-MSE) one-hot the assembler finds little repair work: $+0.4$ pp at em@0 (0.980 $\to$ 0.984, interval crossing zero; clean $+0.5$, ma@0 $+0.5$), and the residual census shows why: its fragmentation-type pairs ($t\to t$, $p\to p$) disappear under stacking while the T-miss pairs ($qrs\to qrs$, $qrs\to p$) survive, the assembler's own failure signature (Appendix A). On the focal convention (Appendix N) the same assembler still finds repair work, $+6.0$ pp at em@0 (0.930 $\to$ 0.989; ma@0 $+5.4$, em@6 $+2.7$, clean $+2.1$), because that convention's residual illegality is fragmentation-type. The stacked one-hot trails the bare field in every audited cell under both conventions (matched-convention gaps: em@0 0.984 vs. 0.994, ma@0 0.990 vs. 0.996, clean 0.998 vs. 0.999; largest focal gap ma@0: 0.983 vs. 0.996) and on F1 (0.845/0.848 vs. 0.861 at em@0); the matched convention's 0.98 tier rests on the selection crutch alone, its single-draw and mean-aggregation collapses (Sec. 4.4) untouched, and the curation trade is the same as on the field (full_cyc $+7.3$ $[6.26, 8.43]$ at em@0 for -0.5 pp F1, T deletions 542 $\to$ 974). On the field's cheap path, one draw plus assembly ($\approx$96 ms) plateaus at 0.937, 5.7 pp below the select tier: selection, not rules, sets the 0.99 tier.

Scope. These cells fix the reading of the replacement claim: generation replaces rules *as the structural-repair mechanism* (a bounded null on the primary metric once structure lives in the representation, on the field and the matched one-hot alike; the focal one-hot's $+6.0$ shows the assembler still repairs where the residual type is fragmentation), while the assembler survives demoted to an optional inter-beat curator whose $+6.6$ pp is bought by deletion at -1 pp F1. A sampling head that reproduces the curation gain without the symbolic step (cycle-aware sampling) would complete the replacement; the measured $+6.6$ pp interval is its target.

P DECODER-SYMMETRY CONTROL

The main decoders are asymmetric by construction: the field decodes per channel with threshold $\tau = 0.15$, connected components, and analytic extrapolation, while the one-hot decodes per point by argmax. A reader could worry that part of the legality gap is bought by the decode convention rather than the target: thresholding with connected components is anti-fragmentary, and extrapolation builds in a shape prior. This control swaps the decoder axis on both sides, everything else fixed: identical 16-candidate pools under the unified seeding protocol (seed 42 on clean/em@0/ma@0; seeds 43/44 at em@0). The arms: the field with its decoder and with the extrapolation removed (pure τ-threshold + connected components, same class order and minimum support); the one-hot with argmax and with a per-class threshold 0.5 + connected components (the decision-boundary threshold, later class winning overlaps, mirroring the field's conventions).

Table 10: Decoder-symmetry control at em@0 (seed 42; legality at the stated operating points, F1 at $K{=}8$ select). Removing the field's extrapolation and re-decoding the one-hot through a matched threshold change no selection-tier reading; thresholds away from the decision boundary ($0.3/0.7$) partially rescue the one-hot mean. Clean and ma@0 behave identically (bit-exact across decoders).

Arm	$K{=}1$	$K{=}8$ select	mean@2	mean@8	F1
Field, threshold + extrapolation	0.888	0.994	0.826	0.843	0.861
Field, pure threshold	0.888	0.994	0.826	0.843	0.861
One-hot, argmax	0.794	0.980	0.000	0.013	0.849
One-hot, threshold 0.5	0.794	0.980	0.000	0.013	0.849
One-hot, threshold 0.3	0.812	0.983	0.753	0.770	0.850
One-hot, threshold 0.7	0.807	0.981	0.778	0.789	0.851

First, the decoder axis does not carry the selection-tier or single-draw gaps (Table 10). Removing the extrapolation leaves every legality and F1 reading bit-identical, in every condition and seed: its role is ms-level boundary recovery (Appendix L), not structure. The decision-boundary per-class threshold *is* the argmax convention: 0.5-threshold decoding reproduces argmax bit-identically (empirically, in every condition and seed; the exact identity needs the per-class readouts to sum to one), and the paired decoder-step contrasts are exactly zero (0.0 $[0.0, 0.0]$ pp on both sides, record-cluster bootstrap). The em@0 representation step is $+1.37$ $[0.48, 2.37]$ under default and symmetric decoders alike (this control's own record-cluster bootstrap, seed 0; the main text's $[0.48, 2.42]$ is the single-cell interval under the per-pair derived RNG and $[0.49, 2.34]$ the joint worst-cell envelope, reconciled in Appendix B: three bootstrap implementations of one estimand); permissive (0.3) and strict (0.7) one-hot thresholds move $K{=}8$ by at most $+0.3$ pp and single draws by $+1.9$ pp, so against the friendliest one-hot decode the step is $+1.07$ pp.

Second, the per-seed disclosure the symmetric protocol makes possible: the per-seed em@0 representation steps are $+1.37$ $[0.47, 2.42]$ (seed 42), $+0.06$ $[-0.56, 0.64]$ (seed 43), and $+1.55$ $[0.59, 2.80]$ (seed 44), record-cluster CIs on this protocol's candidate pools. One seed is a rounding-level tie whose interval crosses zero, the quantitative face of the 0.08 pp band gap of Sec. 4.2; the merged three-seed step, $+0.99$ $[0.39, 1.75]$, excludes zero (Appendix B). The joint interval quoted there is the seed-42 protocol, and the seed-robust statements remain the single-draw step ($+3.2$ to $+11.4$ pp; Sec. 4.4) and the mean operator below.

Third, the mean tier is decode-conditional: the control's honest finding. Under argmax (equivalently 0.5) the one-hot mean@2 is 0.000 at every seed and condition, while per-class thresholds away from the boundary rescue most of it: mean@2 reaches 0.70–0.81 at 0.3 and 0.78–0.88 at 0.7 across seeds (at seed 43 a 0.7-threshold one-hot mean, 0.883, exceeds the field's 0.836). The field's mean at its native $\tau = 0.15$ stays 0.83–0.84, and a dedicated τ sweep of the mean aggregation (em@0 and clean, $\tau \in \{0.05, 0.15, 0.25, 0.35\}$) shows it never collapses at any tested threshold: mean@2 varies by at most 3 pp (0.798–0.828 at em@0; 0.938–0.943 clean), and mean@8 ranges 0.704–0.907, rising monotonically with τ (permissive thresholds admit disagreement fragments). The surviving asymmetry is thus one of floors, not of flatness: the one-hot mean is exactly 0.000 at the decision boundary that every established convention (argmax; the DiffAct deterministic readout) uses, at every tested loss, sampler, and seed (Appendix N), whereas the field mean never falls below 0.70 anywhere in the sweep and averages legally with no threshold search. The averagability statement of Sec. 4.4 is

accordingly scoped to decision-boundary decoding rather than claimed as invariance to permissive decoding that no default convention employs.

Q OUT-OF-DOMAIN PROBE: HEART-SOUND SEGMENTATION

A single-seed, two-condition pilot asks whether the operating-point conclusions leave the ECG domain, reusing the ECG protocol unchanged (same 1.77M U-Net, in-model corruption, DDIM budget, $K{=}16$ candidates; no CIs, no seed replication: probe status). Data: PhysioNet/CinC 2016 hand-corrected four-state labels (Liu et al., 2016) (Springer-algorithm segmentations with manual correction (Springer et al., 2016); S1/systole/S2; diastole maps to background), 947 recordings split 686/121/140 into 6,595/494/725 windows at 500 Hz; the grammar is the fixed cycle $S1{\to}sys{\to}S2{\to}S1$, checked as the adjacent-run transition set (diastole is background; ground-truth legality 1.0 on all splits); evaluation covers clean and 0 dB.

Table 11: Heart-sound transfer probe (test legality and macro F1; run ratio = predicted/ground-truth run count, median over windows).

	clean		0 dB	
Metric	one-hot	field	one-hot	field
legality select@1	0.927	**0.978**	0.808	**0.956**
legality select@8	1.000	1.000	1.000	1.000
legality mean@2	0.793	**0.877**	0.520	**0.745**
legality mean@16	**0.920**	0.803	**0.763**	0.590
F1 single draw	0.829	**0.846**	**0.714**	0.685
F1 mean@2	0.837	**0.853**	0.706	**0.724**
run ratio mean@2	1.067	**1.000**	1.375	**1.000**

Three readings (Table 11). **First, the one-hot mean collapse reproduces out of domain**: 0.808 $\to$ 0.520 at $K{=}2$ under 0 dB, with the fragmentation signature quantified (run ratio 1.375 vs. the field's 1.000). **Second, the operating points transfer with a saturation caveat**: the mean@2 conclusions replicate (one-hot collapses, the field does not; F1 and run integrity in step; onset hit rates within $\pm$50 ms consistently higher, $+0.02$ to $+0.05$), while select@8 saturates at 1.000 for both arms, leaving the ECG ordering at that tier untestable here; single-draw F1 under noise is the one metric where the arms trade places (0.714 vs. 0.685). The transfer is cross-modality on the same organ (acoustic/mechanical vs. electrical). **Third, the regime boundary**: field means degrade beyond $K{=}2$ (0.745 $\to$ 0.590 at 0 dB, where one-hot overtakes at 0.763), unlike ECG, where they are stable at every measured K. The decoding geometry marks the regime: where silent stretches and murmurs leave the posterior cross-class multimodal, averaging leaves a minority-mode bump whose height tracks its mode mass; it survives the default $\tau = 0.15$ whenever that mass exceeds τ and decodes as a spurious cross-class run, so large-K averaging needs τ above the minority mode's mass. Averagability is thus a joint property of the representation, the posterior's modal structure, τ, and the task; the ECG claims of Sec. 4.4 stand within their measured regime. Onset MAE is excluded (cycle-slip contaminated by windows starting at arbitrary phase; hit rates are unaffected).