# CODEBOOK-ANCHORED PREDICTION ACROSS PHYSIOLOGICAL VIEWS

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

## ABSTRACT

Physiological sensors capture related but non-equivalent views of the same cardiovascular interval. Cross-view reconstruction is difficult because the inverse mapping is ambiguous and continuous latent spaces provide no stable correspondence across sensors. We introduce PHYSIOANCHOR, a codebook-anchored predictor that learns a shared four-level residual-vector-quantized (RVQ) tokenizer and freezes its latent vocabulary during cross-view prediction. Modality-specific encoders and decoders retain view-specific signal structure, while nearest-codeword projection constrains predicted target latents before waveform decoding. We use a fixed subject-disjoint split of MIMIC-IV Waveform v0.1.0 with 138/20/40 train/validation/test subjects. Against the strongest evaluated external baseline, PHYSIOANCHOR reduces subject-level PPG-to-ECG RMSE from $0.526$ to $0.493$ and increases R-peak F1 from $0.704$ to $0.739$. Under matched codebook parameter capacity, four-level RVQ reduces tokenizer RMSE from $0.171$ to $0.070$ relative to flat VQ. Using the same predictor checkpoint, nearest-codeword projection improves pooled-window F1 from $0.792$ to $0.827$. The same interface supports 25-to-250/500-Hz ECG reconstruction with R-peak F1 of $0.918/0.912$. In conclusion, these results show that our two-stage architecture supports both cross-sensor and cross-rate physiological reconstruction through a shared, frozen RVQ latent space within the evaluated MIMIC-IV cohort.

## 1 INTRODUCTION

Predictive representation learning infers target features from observed context. In I-JEPA and V-JEPA, context and target observations come from the same visual modality and are encoded within a common representation family (Assran et al., 2023; Bardes et al., 2024). Our setting instead connects heterogeneous physiological views through view-specific networks. ECG and PPG arise from different sensing processes, whereas rate-limited ECG retains only part of the target bandwidth. Even when these views cover the same cardiovascular interval, their continuous features are not guaranteed to be directly comparable. Cross-view prediction therefore requires an explicit source-to-target latent transformation.

We structure this transformation with a shared four-level RVQ tokenizer. During Stage 1, view-specific encoders and decoders are trained against the same codebooks, which provide common quantization anchors without requiring identical token use across views. During Stage 2, separate predictors are trained for PPG-to-ECG and cross-rate ECG reconstruction against frozen target-encoder outputs. Nearest-codeword projection then constrains each predicted target to the learned RVQ vocabulary before waveform decoding. Thus, the predictors learn the task-specific cross-view mappings, while the shared codebooks define and regularize their prediction interface.

Built on this interface, PHYSIOANCHOR performs codebook-anchored prediction across sensor modalities and sampling rates. A compact Transformer predicts ECG latents from PPG features and five static covariates, while a masked latent operator predicts high-rate ECG latents from low-rate ECG. The two predictors are trained separately but reuse the frozen RVQ vocabulary, target-side projection, and ECG decoding interface. This design preserves task-specific transformations while exposing both tasks to a common discrete target vocabulary.

Our contributions are:



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.

- We formulate cross-view physiological reconstruction as prediction into a frozen, shared discrete reference frame.
- We combine a four-level RVQ tokenizer, modality-specific renderers, static-context prompting, latent prediction, and target-side nearest-codeword projection.
- We evaluate the design with subject-disjoint baselines, matched tokenizer and adaptation controls, 250/500-Hz ECG reconstruction.

## 2 Related work

**Predictive representation learning.** I-JEPA predicts representations of masked image regions from visible context, and V-JEPA extends feature prediction to spatiotemporal video regions (Assran et al., 2023; Bardes et al., 2024). Both methods construct context and target representations within a common visual feature space, avoiding direct prediction of low-level pixels. Extending this principle to physiological signals is nontrivial because sensor modalities and sampling rates define heterogeneous views that may require view-specific representations.

**Discrete latent representations.** Vector-quantized autoencoders map continuous features to a learned discrete vocabulary (van den Oord et al., 2017). Residual vector quantization represents an input through successive codewords and has enabled compact reconstruction of structured signals in neural audio codecs (Zeghidour et al., 2022; Défossez et al., 2023). This literature primarily uses discrete vocabularies for within-modality representation, compression, or generation. Their role as prediction targets across heterogeneous physiological views remains less established.

**Physiological signal translation and reconstruction.** Prior PPG-to-ECG methods reconstruct ECG directly, use adversarial or diffusion objectives, or translate through continuous latent spaces. UA-P2E models conversion uncertainty with diffusion (Belhasin et al., 2024), whereas PPGFlowECG aligns ECG and PPG representations before latent rectified-flow generation (Fang et al., 2025). These approaches address the ambiguity of cross-sensor reconstruction through generative modeling or continuous latent alignment. Cross-rate ECG reconstruction poses a related representation problem because rate-limited inputs omit part of the target signal bandwidth. Together, these settings motivate investigating whether a shared discrete prediction target can support both sensor and rate transformations.

## 3 Method

### 3.1 Problem formulation: multi-view physiological state

Let $x^{(m,r)} \in \mathbb{R}^T$ denote an 8-s waveform from modality $m \in \{\mathrm{ECG}, \mathrm{PPG}\}$ and sampling rate $r$. Let $s \in \mathbb{R}^5$ contain age, sex, height, weight, and body mass index. We seek

$$\hat{x}^{(b)} = D_b\Big(Q\Big(F_{a\to b}(E_a(x^{(a)}), s)\Big)\Big), \tag{1}$$

where $E_a$ is the source-view encoder, $Q$ is the shared four-level RVQ tokenizer, $D_b$ is the target-view decoder, and $F_{a\to b}$ is a cross-view latent predictor. We evaluate PPG-to-ECG and low-rate-to-high-rate ECG predictors. This factorization fixes the target coordinate system while each cross-view mapping is learned.

### 3.2 Stage 1: shared discrete anchor space

Each modality-specific one-dimensional encoder maps its input to $Z^{(m)} = E_m(x^{(m)}) \in \mathbb{R}^{N\times d}$ with $d = 64$. A prompt network

$$p(s) = W_2\,\mathrm{GELU}(W_1 s + b_1) + b_2 \in \mathbb{R}^{64} \tag{2}$$

is broadcast over time, yielding $\widetilde{Z}^{(m)} = Z^{(m)} + p(s)$. This prompt supplies static subject context that may not be recoverable from a short waveform.

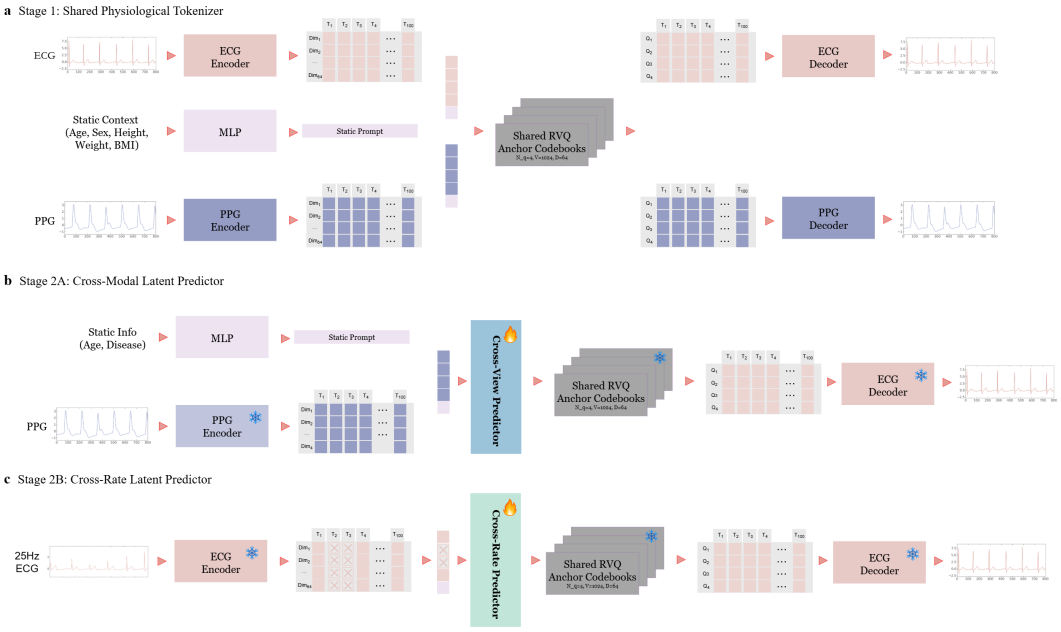



Figure 1: Overview of PHYSIOANCHOR. Stage 1 trains modality-specific encoders and decoders around a shared four-level RVQ vocabulary. Stage 2 freezes this interface and learns either a PPG-to-ECG latent predictor or a masked cross-rate predictor. Both routes project predicted features through the frozen codebooks before ECG decoding. Fire and snowflake symbols denote trainable and frozen components, respectively.

The four-level RVQ contains codebooks $\mathcal{C}_\ell = \{e_{\ell,k}\}_{k=1}^{1024}$, $e_{\ell,k} \in \mathbb{R}^{64}$. With $R_0 = \widetilde{Z}$, cosine nearest-neighbor lookup and residual recursion give

$$q_\ell = Q_\ell(R_{\ell-1}), \qquad R_\ell = R_{\ell-1} - q_\ell, \qquad Z_q = \sum_{\ell=1}^{4} q_\ell. \tag{3}$$

Codebooks use 10-iteration $k$-means initialization and a commitment weight of 1.0. Sharing $Q$ gives every view access to the same anchors without requiring identical token use. Modality-specific decoders $D_m$ map quantized latents back to ECG or PPG. The reconstruction objective combines sample, angular, and first-difference fidelity:

$$\mathcal{L}_{\text{rec}} = \|x - \hat{x}\|_1 + 0.5\big(1 - \cos(x, \hat{x})\big) + 0.5\|\nabla x - \nabla \hat{x}\|_1, \tag{4}$$

and $\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{VQ}}$.

### 3.3 STAGE 2: FROZEN-SPACE CROSS-VIEW PREDICTION

After Stage 1, the modality-specific encoders, codebooks, decoders, and their normalization statistics are frozen. For PPG-to-ECG translation, a Transformer receives a prompt token and PPG latent sequence,

$$H = \text{Transformer}([p(s); E_{\text{PPG}}(x^{\text{PPG}})]), \qquad \widehat{Z}^{\text{ECG}} = g(H_{2:}), \tag{5}$$

where $g : \mathbb{R}^{64} \to \mathbb{R}^{64}$. It minimizes continuous target-latent MSE against the frozen ECG encoder output. At inference, either $D_{\text{ECG}}(\widehat{Z})$ is decoded directly or $\widehat{Z}$ is projected through the frozen RVQ as in Eq. 1. The same-checkpoint comparison in Section 4.4 isolates nearest-codeword projection.

Writing the frozen target representation as $Z_b^\star = \text{sg}[E_b(x^{(b)})]$, the prediction objective is

$$\mathcal{L}_{\text{pred}} = \frac{1}{N} \sum_{t=1}^{N} \|\widehat{Z}_t^{(b)} - Z_{b,t}^\star\|_2^2, \tag{6}$$

where $\mathrm{sg}$ denotes stop-gradient. Freezing the target stack makes this coordinate system stationary. Projecting $\widehat{Z}^{(b)}$ through $Q$ further constrains decoding to sums of learned codewords.

For cross-rate reconstruction, a masked latent operator observes the 25-Hz ECG representation and predicts the high-rate latent sequence. Both tasks reuse the tokenizer and target decoder, although their predictors are trained separately.

### 3.4 Optimization and capacity

All configurations use AdamW with a learning rate of $10^{-3}$, weight decay of $10^{-4}$, and a batch size of 64 (Loshchilov & Hutter, 2019). They follow the same subject split, random seed, optimization budget, and checkpoint-selection rule, with one training run per configuration. The full architecture contains 90M parameters. For the flat-VQ comparison, both tokenizers are trained for 50 epochs under matched codebook capacity (262,144 scalars), using either one $4096 \times 64$ codebook or four $1024 \times 64$ codebooks. This comparison isolates quantizer structure at a fixed codebook parameter count; bitrate is not matched.

## 4 Experiments

### 4.1 Dataset, split, and evaluation protocol

We use MIMIC-IV Waveform Database v0.1.0, which contains 200 records from 198 subjects and links to MIMIC-IV Clinical Database v2.0 (Moody et al., 2022; Johnson et al., 2022). We fix a subject-disjoint split of 138/20/40 train/validation/test subjects. Both stages use this split. Training subjects may overlap across stages, but the 40 test subjects are excluded from all training, validation, and checkpoint selection. Adjacent or overlapping windows never cross splits.

Each observation is an 8-s paired ECG–PPG window. ECG and PPG are resampled to 100 Hz, and every comparison receives the same fixed preprocessing. We create 25-Hz ECG inputs using predefined low-pass anti-alias filtering followed by downsampling rather than sample slicing. No reported comparison uses post-hoc temporal alignment.

**Metrics and statistical unit.** The primary cross-modal table pools windows within each test subject, computes one metric per subject, and reports the mean$\pm$SD across 40 subjects. These SDs quantify between-subject heterogeneity, not retraining variability. We detect R peaks in real and generated ECG using the same NeuroKit2 `ecg_peaks(method="neurokit")` configuration (Makowski et al., 2021). Hungarian assignment matches predicted and reference peaks one-to-one within $\pm 50$ ms. TP, FP, and FN are pooled across each subject's windows before computing F1. Predictions in reference-peak-free windows count as false positives.

Fréchet distance (FD) uses a fixed 32-dimensional representation. We compute Welch PSD over 0.5–40 Hz, normalize total power, apply $\log(\mathrm{PSD} + 10^{-12})$, and standardize using real training-set ECG statistics (Welch, 1967). We then compute one FD from each subject's real and generated window distributions. No time point or window is treated as an independent subject.

### 4.2 Cross-modal prediction and reconstruction

Table 1 compares end-to-end P2E, PPGFlowECG (Fang et al., 2025), UA-P2E (Belhasin et al., 2024), and PhysioAnchor under the same split, preprocessing, and evaluation code. PhysioAnchor achieves lower mean RMSE and higher R-peak F1 than UA-P2E, the strongest evaluated external baseline, with $0.493$ versus $0.526$ and $0.739$ versus $0.704$, respectively. MAE and FD show the same ordering.

Figure 2 visualizes a representative 8-s test window. In this example, PhysioAnchor follows the measured ECG peak timing and waveform morphology more closely than end-to-end P2E, consistent with the subject-level RMSE and R-peak F1 trends in Table 1.

Table 1: PPG-to-ECG reconstruction on 40 test subjects. Values are the mean±SD across subjects. Lower is better for RMSE, MAE, and FD; higher is better for F1.

| Method | RMSE↓ | MAE↓ | R-peak F1↑ | FD↓ |
|---|---|---|---|---|
| P2E | .698±.128 | .489±.104 | .351±.158 | 39.1±11.2 |
| PPGFlowECG | .579±.113 | .377±.093 | .638±.126 | 29.2±8.8 |
| UA-P2E | .526±.103 | .339±.085 | .704±.111 | 24.5±7.6 |
| PHYSIOANCHOR | **.493±.099** | **.319±.080** | **.739±.103** | **22.8±7.1** |

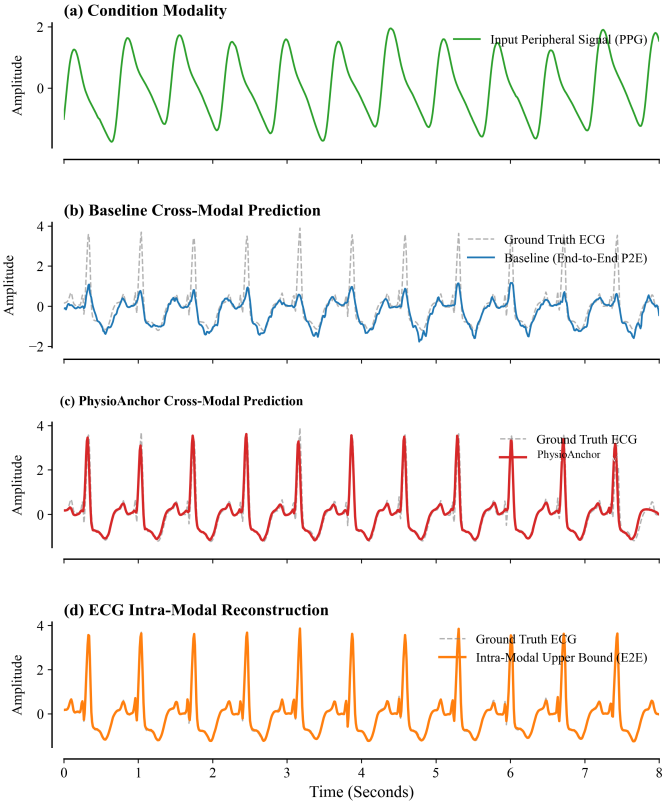



Figure 2: Representative PPG-to-ECG reconstruction over an 8-s window. (a) Input PPG. (b) End-to-end P2E reconstruction against the measured ECG. (c) PHYSIOANCHOR reconstruction from PPG. (d) ECG autoencoding as an intra-modal reference. This example visualizes waveform morphology; Table 1 provides the subject-level comparison.

### 4.3 CROSS-RATE PREDICTION AND RECONSTRUCTION

In the pooled-window evaluation, 25-to-100-Hz reconstruction achieves RMSE 0.0847 and R-peak F1 0.9248, outperforming linear, cubic, and convolutional baselines (Appendix D). Figure 3 shows one corresponding waveform, enlarged QRS complexes, and power spectrum. We also evaluate 25-to-250/500-Hz reconstruction using subject-disjoint data. Table 2 shows similar R-peak F1 at both target rates, while QRS, QT, and QTc errors are higher at 500 Hz. These metrics assess interval and spectral fidelity but do not establish diagnostic equivalence.

Table 2: High-rate ECG reconstruction from 25-Hz input. Values are the mean±SD across test subjects; interval errors are in milliseconds.

| Target | QRS MAE↓ | QT MAE↓ | QTc MAE↓ | F1↑ | PSD-RMSE↓ |
|---|---|---|---|---|---|
| 250 Hz | 9.8±.7 | 14.6±1.2 | 16.1±1.4 | .918±.006 | .0016±.0001 |
| 500 Hz | 10.7±.8 | 15.8±1.3 | 17.3±1.5 | .912±.007 | .0019±.0001 |

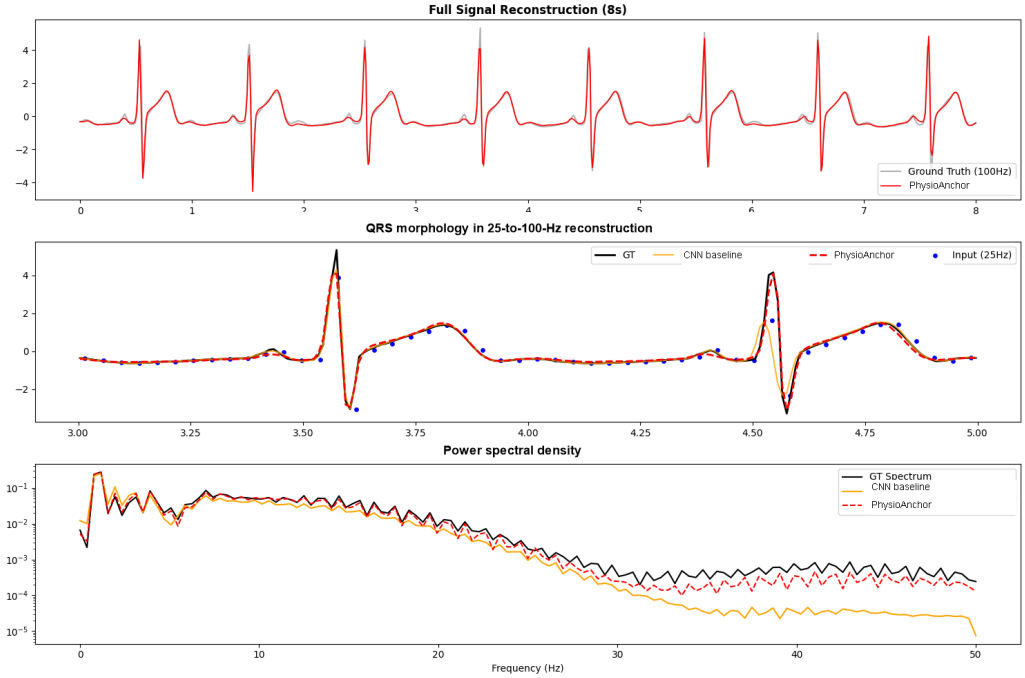



Figure 3: Representative 25-to-100-Hz ECG reconstruction under the pooled-window protocol. The panels show the full 8-s waveform, two enlarged QRS complexes with 25-Hz input samples, and the corresponding power spectra. The CNN baseline attenuates more high-frequency energy in this example, whereas PHYSIOANCHOR more closely follows the measured waveform and spectrum. Table 2 reports subject-level 250/500-Hz interval and spectral results.

### 4.4 CONTROLLED ANALYSES OF THE DISCRETE INTERFACE

**Residual quantization under matched codebook capacity.** We compare flat VQ and four-level RVQ using identical encoder and decoder architectures, latent dimensionality, training objectives, and optimizer-update budgets. Both tokenizers are trained for 50 epochs with 262,144 codebook parameters, using one $4096 \times 64$ codebook or four $1024 \times 64$ codebooks. This comparison matches codebook parameter capacity rather than bitrate. Four-level RVQ improves Stage-1 ECG reconstruction across all four metrics (Table 3). Mean subject-level RMSE decreases from 0.171 to 0.070, a 59.1% relative reduction, while R-peak F1 increases from 0.902 to 0.964. These results support residual quantization for ECG reconstruction under the matched codebook parameter budget.

**Nearest-codeword projection.** We isolate inference-time projection using the same Stage-2 predictor checkpoint and frozen target stack. The two configurations differ only in whether predicted latents undergo nearest-codeword projection before waveform decoding. Projection reduces mean subject-level RMSE from 0.519 to 0.493, a 5.0% relative reduction, and increases R-peak F1 from 0.704 to 0.739 (Table 3). MAE and FD also decrease, showing consistent improvements across the evaluated metrics. In the abnormal-morphology subset analysis, R-peak F1 increases from 0.684 to 0.741. These results support projecting predicted latents onto the learned RVQ vocabulary before ECG decoding.

Table 3: Controlled component studies. Values are mean±SD across the same 40 held-out subjects, with one training run per training configuration. SD describes between-subject variation. Metrics are computed separately for each subject and then averaged with equal subject weights. FD compares each subject's real and reconstructed window-feature distributions. The tokenizer block evaluates Stage-1 ECG reconstruction; the projection and adaptation blocks evaluate PPG-to-ECG reconstruction. The nearest-codeword and frozen rows share the full-model reference reported in Table 1. Bold indicates the best mean within each block.

| Block | Variant | RMSE↓ | MAE↓ | R-peak F1↑ | FD↓ |
|---|---|---|---|---|---|
| Tokenizer | Flat VQ | $0.171 \pm 0.006$ | $0.112 \pm 0.005$ | $0.902 \pm 0.009$ | $27.95 \pm 6.7$ |
| | 4-level RVQ | $\mathbf{0.070 \pm 0.003}$ | $\mathbf{0.043 \pm 0.002}$ | $\mathbf{0.964 \pm 0.004}$ | $\mathbf{16.2 \pm 6.5}$ |
| Projection | Continuous | $0.519 \pm 0.102$ | $0.337 \pm 0.082$ | $0.704 \pm 0.106$ | $25.3 \pm 7.3$ |
| | Nearest codeword | $\mathbf{0.493 \pm 0.099}$ | $\mathbf{0.319 \pm 0.080}$ | $\mathbf{0.739 \pm 0.103}$ | $\mathbf{22.8 \pm 7.1}$ |
| Adaptation | Frozen | $\mathbf{0.493 \pm 0.099}$ | $\mathbf{0.319 \pm 0.080}$ | $\mathbf{0.739 \pm 0.103}$ | $\mathbf{22.8 \pm 7.1}$ |
| | Last quantizer | $0.506 \pm 0.100$ | $0.329 \pm 0.081$ | $0.728 \pm 0.104$ | $23.6 \pm 7.2$ |
| | Decoder + last quantizer | $0.524 \pm 0.103$ | $0.342 \pm 0.083$ | $0.713 \pm 0.107$ | $25.0 \pm 7.4$ |

**Freezing the target stack.** We compare a fully frozen target stack with two partial-adaptation strategies under the same training budget. These strategies unfreeze either the last quantizer alone or both the decoder and the last quantizer. The frozen configuration achieves the best mean values across all four metrics (Table 3). Mean subject-level RMSE is 0.493 with the frozen stack, compared with 0.506 and 0.524 for the two partial-adaptation strategies. These results favor retaining the pretrained target stack during cross-view predictor training under the evaluated protocol.

**Codebook utilization and organization.** Across RVQ levels 1–4, ECG perplexity increases from 622 to 705, and PPG perplexity increases from 604 to 668. Over the same levels, dead-code rates decrease from 8.2% to 6.1% for ECG and from 10.1% to 8.0% for PPG. Layer-wise probes associate the incremental inclusion of levels 1–3 with error reductions in the 0.5–5, 5–15, and 15–40 Hz bands, respectively. The reductions are 58% (subject-bootstrap 95% CI, 52–64%), 43% (37–49%), and 31% (24–37%). Adding level 4 reduces QRS-width error by 18% (95% CI, 11–24%). These probes link successive RVQ levels to improvements in spectral reconstruction and local ECG morphology.

**Static context.** A separate pooled-window analysis evaluates static conditioning with quantization retained. Removing static conditioning decreases R-peak F1 from 0.8265 to 0.8032. RMSE increases from 0.4245 to 0.7127, and FD increases from 16.18 to 33.77. Within this analysis, static conditioning improves both waveform reconstruction and R-peak detection. These results support retaining static context in the evaluated reconstruction pipeline.

## 5 Discussion

The results establish a discrete prediction interface for heterogeneous physiological views. In PPG-to-ECG reconstruction, PhysioAnchor achieves lower subject-level RMSE and higher R-peak F1 than the strongest evaluated external baseline. The same-checkpoint control further shows that nearest-codeword projection improves reconstruction across all four evaluated metrics. The cross-rate experiments extend the interface to low-rate and high-rate ECG views using a separate task-specific predictor.

The tokenizer experiments clarify the role of residual quantization. Under matched codebook parameter capacity, four-level RVQ improves Stage-1 reconstruction over flat VQ across RMSE, MAE, R-peak F1, and FD. Increasing perplexity and decreasing dead-code rates indicate that successive residual levels remain actively used. The frequency-band and QRS-width probes further associate these levels with progressive reconstruction refinement. These findings support the residual structure of the tokenizer without assigning exclusive physiological semantics to individual codebooks.

The shared tokenizer gives the view-specific encoders access to a common codeword inventory and supplies a fixed target vocabulary for cross-view prediction. The present controls compare flat and residual quantization but do not compare shared codebooks with independently learned modality-

specific codebooks. The results therefore support the complete codebook-anchored interface, while the causal contribution of codebook sharing remains to be isolated.

PHYSIOANCHOR extends feature prediction from contexts and targets within one representation family to physiological views produced by different sensors or sampling rates. The PPG-to-ECG and cross-rate operators remain task-specific, while the RVQ vocabulary, projection mechanism, and target decoding interface are reused. This structure could provide an observation interface for future temporal or contrastive objectives, although the present experiments evaluate contemporaneous reconstruction rather than representation transfer or temporal rollout.

**When reconstruction is appropriate.** Cross-view reconstruction can be useful when wearable PPG is available but an ECG-like signal is required by an existing signal-processing workflow. Cross-rate reconstruction is also relevant when acquisition, storage, or transmission preserves only low-rate ECG. Directly measured ECG remains preferable whenever it is available. For a fixed clinical endpoint, direct prediction from the observed signal may also be more appropriate than waveform reconstruction. The present experiments evaluate signal fidelity and do not establish a downstream diagnostic advantage.

**Limitations.** All experiments use one fixed MIMIC-IV cohort of 198 subjects and one training run per configuration. The reported SDs quantify variation across test subjects rather than sensitivity to model initialization. The available evaluation also lacks demographic, pathology, and acquisition-device strata, limiting conclusions about subgroup and cross-domain generalization.

Several methodological questions remain open. The experiments do not isolate shared codebooks against modality-specific alternatives. PPG-to-ECG reconstruction is intrinsically ambiguous, and nearest-codeword projection may produce physiologically plausible but individually incorrect morphology. The static-context ablation also shows that reconstruction quality depends materially on the available metadata. Finally, the high-rate experiments measure interval, peak, and spectral fidelity without evaluating diagnostic equivalence to measured high-rate ECG.

## 6 CONCLUSION

PHYSIOANCHOR formulates heterogeneous physiological reconstruction as prediction into a shared, frozen discrete vocabulary. On subject-disjoint MIMIC-IV data, a four-level RVQ tokenizer, task-specific latent predictors, and nearest-codeword projection support both PPG-to-ECG and cross-rate ECG reconstruction. Matched component controls support residual quantization, target-side projection, and retention of the frozen target stack under the evaluated protocol. These results establish a codebook-anchored interface for cross-view physiological reconstruction within the studied cohort.

### ETHICS STATEMENT

The study uses the deidentified, credentialed MIMIC-IV Waveform and Clinical databases. The source dataset reports institutional review-board approval and a waiver of individual consent following deidentification of protected health information. This study enrolled no new participants and releases no patient-level data. Reconstructed ECG may contain plausible but incorrect morphology and should not replace directly measured clinical ECG, particularly for acute diagnosis. The absence of demographic and pathology strata also precludes subgroup-level assessment of prospective clinical use.

### REPRODUCIBILITY STATEMENT

Section 4 and Appendix B specify the dataset versions, subject counts, data split, window construction, anti-aliasing, metric implementations, aggregation procedures, and statistical units. Appendix A records the model and training hyperparameters. Appendix E documents the matched tokenizer control, codebook utilization statistics, and layer-wise functional probes. The training and evaluation code is not included in the anonymous submission and is planned for public release upon publication.

## AI USE STATEMENT

Generative AI tools, including OpenAI Codex, were used to improve the language, organization, clarity, and presentation of the manuscript. They were not used to generate data, labels, experimental measurements, or statistical results.

## REFERENCES


Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Omer Belhasin, Idan Kligvasser, George Leifman, Regev Cohen, Erin Rainaldi, Li-Fang Cheng, Nishant Verma, Paul Varghese, Ehud Rivlin, and Michael Elad. Uncertainty-aware ppg-2-ecg for enhanced cardiovascular diagnosis using diffusion models, 2024. URL https://arxiv.org/abs/2405.11566. arXiv:2405.11566.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=ivCd8z8zR2.

Xiaocheng Fang, Jiarui Jin, Haoyu Wang, Che Liu, Jieyi Cai, Yujie Xiao, Guangkun Nie, Bo Liu, Shun Huang, Hongyan Li, and Shenda Hong. PPGFlowECG: Latent rectified flow with cross-modal encoding for ppg-guided ecg generation and cardiovascular disease detection, 2025. URL https://arxiv.org/abs/2509.19774. arXiv:2509.19774.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, 2022. doi: 10.13026/7vcr-e114. Version 2.0.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Dominique Makowski, Tam Pham, Zen J Lau, Jan C Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S H Annabel Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53:1689–1696, 2021. doi: 10.3758/s13428-020-01516-y.

Benjamin Moody, Sicheng Hao, Brian Gow, Tom Pollard, Wei Zong, and Roger Mark. MIMIC-IV Waveform Database. *PhysioNet*, 2022. doi: 10.13026/a2mw-f949. Version 0.1.0.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Peter Welch. The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967. doi: 10.1109/TAU.1967.1161901.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2022. doi: 10.1109/TASLP.2021.3129994.

## A Implementation details

Each modality-specific encoder outputs token vectors with dimension $d = 64$. The prompt MLP has dimensions $5 \to 64 \to 64$ and uses GELU. Each of the four shared RVQ codebooks contains 1,024 vectors with dimension 64. Codeword lookup uses cosine distance, $k$-means initialization runs for 10 iterations, and the commitment weight is 1.0.

Stage 1 broadcasts the prompt over token time. Stage 2 prepends the prompt as a Transformer token and applies a $64 \to 64$ output projection after removing the prompt position. Models use AdamW with learning rate $10^{-3}$, weight decay $10^{-4}$, and batch size 64. Compared methods use the same documented checkpoint-selection rule. Both encoders, all codebooks and decoders, and their normalization statistics are frozen in the primary Stage-2 configuration.

## B Evaluation protocol details

**Subject separation.** The 138/20/40 split is applied before window generation. Every window from a subject belongs to one split, so adjacent or overlapping windows cannot cross split boundaries. Tokenizer and predictor checkpoint selection use only the training and validation subjects.

**R-peak F1.** The same NeuroKit2 detector and parameters are applied to reference and reconstructed ECG in each 8-s window. Hungarian matching enforces one-to-one peak pairs within $\pm 50$ ms. TP, FP, and FN are pooled across the windows belonging to each subject before computing $2\mathrm{TP}/(2\mathrm{TP}+\mathrm{FP}+\mathrm{FN})$. Predicted peaks in a reference-peak-free window count as false positives. The subject-level tables average the resulting values equally across the 40 held-out subjects.

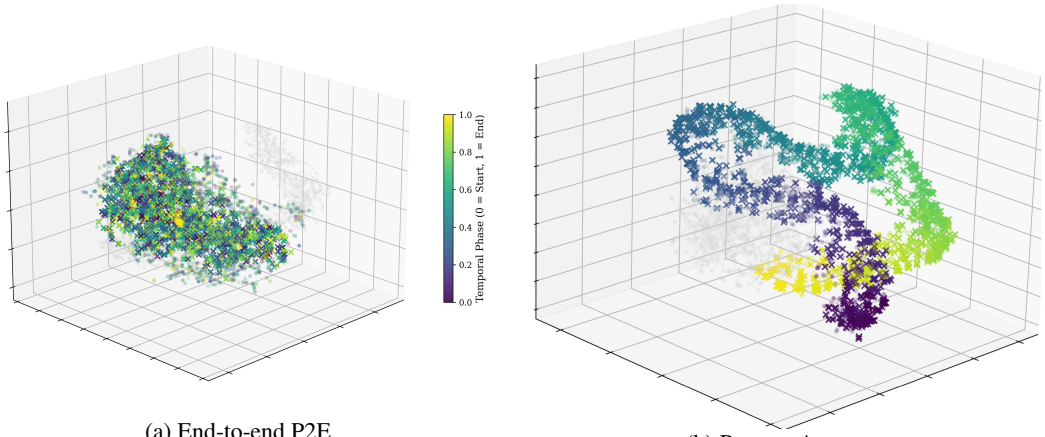
**Fréchet distance.** Each window is transformed into 32 log-normalized Welch-PSD features over 0.5–40 Hz. Total spectral power is normalized before applying $\log(\mathrm{PSD} + 10^{-12})$. Feature standardization uses statistics fitted only on real training-set ECG.

For the subject-level analyses, FD is calculated separately between each subject's real and reconstructed window-feature distributions. The reported value is then averaged equally across the 40 test subjects. The pooled-window analyses in Appendix D instead compare feature distributions after pooling all evaluation windows. Values from the subject-level and pooled-window protocols therefore represent different aggregation units and should not be compared directly.

**Statistical reporting.** Mean$\pm$SD in Tables 1 and 3 quantifies variation across the same 40 held-out subjects for predictions from one trained checkpoint per configuration. These SDs do not quantify retraining variability. Confidence intervals for the layer-wise functional probes are obtained by subject-level bootstrap resampling. The external-baseline, projection, and partial-adaptation comparisons are reported descriptively without formal hypothesis tests. Independent multi-seed retraining is not included in the present evaluation.

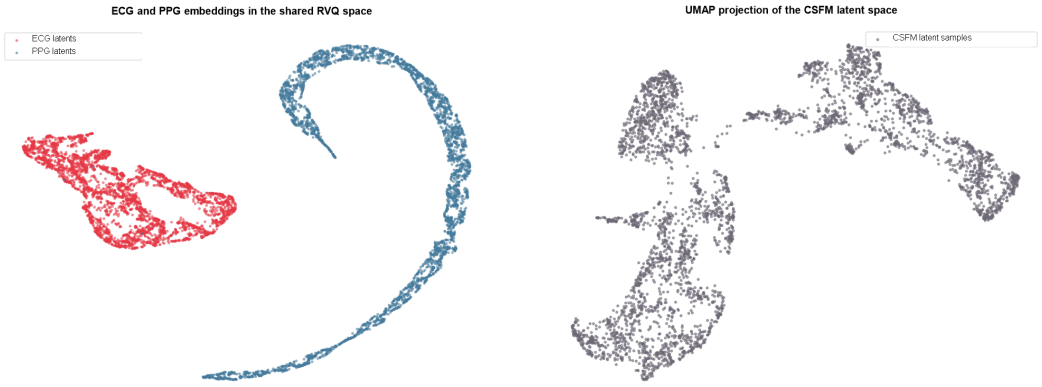
## C Exploratory latent-space visualizations

We project latent samples using t-SNE and UMAP as exploratory diagnostics. These visualizations illustrate the organization of the learned representations but are not used as quantitative evidence of identifiable geometry or cross-modal alignment.




(a) End-to-end P2E (b) PHYSIOANCHOR

Figure 4: Exploratory 3D t-SNE projections of cross-modal latent prediction. Gray points denote input PPG embeddings, while target and predicted ECG embeddings are colored by normalized cardiac-cycle phase. The projection for PHYSIOANCHOR shows more visible phase organization than the P2E projection in this run. Because t-SNE is nonlinear and sensitive to initialization, the visualization does not establish global manifold geometry or causal alignment.




(a) Four-level RVQ tokenizer (b) CSFM comparison

Figure 5: Exploratory UMAP projections of the learned representations. Panel (a) displays ECG and PPG embeddings from the shared four-level RVQ tokenizer, and panel (b) displays the CSFM comparison. Apparent separation or overlap is projection-dependent and does not quantify cross-modal alignment by itself. Appendix E reports the quantitative codebook utilization and layer-wise functional probes.

## D POOLED-WINDOW EVALUATIONS

Table 4: Pooled-window PPG-to-ECG reconstruction from one trained checkpoint per method. Table 1 provides the primary subject-level comparison.

| Method | RMSE↓ | MAE↓ | F1↑ | FD↓ |
|---|---|---|---|---|
| P2E | .6805 | .4760 | .3724 | 37.4933 |
| PHYSIOANCHOR | .4245 | .2517 | .8265 | 16.1779 |
| PHYSIOANCHOR without static context | .7127 | .4717 | .8032 | 33.7659 |

Table 5: Pooled-window 25-to-100-Hz ECG reconstruction.

| Method | RMSE↓ | MAE↓ | PCC↑ | F1↑ | PSD-RMSE↓ | FD↓ |
|---|---|---|---|---|---|---|
| Linear | .3708 | .1634 | .9212 | .7231 | .0138 | 11.7139 |
| Cubic | .4510 | .1910 | .8862 | .7529 | .0078 | 15.0424 |
| CNN | .2053 | .1028 | .9773 | .8349 | .0039 | 1.3209 |
| PHYSIOANCHOR | **.0847** | **.0353** | **.9956** | **.9248** | **.0013** | **.0406** |

## E MATCHED CONTROLS AND CODEBOOK DIAGNOSTICS

**Capacity match.** The flat tokenizer uses one $4096 \times 64$ codebook, whereas the RVQ tokenizer uses four $1024 \times 64$ codebooks. Both configurations contain 262,144 learned codebook scalars. We hold the encoder, decoder, latent dimension, losses, optimizer updates, initialization policy, and 50-epoch training budget fixed. One flat index and four residual indices have different coding costs, so this control matches codebook parameter capacity rather than bitrate.

Table 6: Tokenizer reconstruction on the 40 held-out subjects. Values are mean±SD across subjects. PSD error provides an additional spectral reconstruction measure beyond the metrics reported in Table 3.

| Tokenizer | RMSE↓ | R-peak F1↑ | PSD error↓ | FD↓ |
|---|---|---|---|---|
| Flat VQ | $.171 \pm .006$ | $.902 \pm .009$ | $.000462 \pm .000024$ | $27.95 \pm 6.7$ |
| 4-level RVQ | $\mathbf{.070 \pm .003}$ | $\mathbf{.964 \pm .004}$ | $\mathbf{.000177 \pm .000012}$ | $\mathbf{16.2 \pm 6.5}$ |

Table 7: Layer-wise codebook utilization and functional probes. Intervals are subject-bootstrap 95% CIs for the relative error reductions.

| Layer | ECG ppl. | PPG ppl. | ECG dead | PPG dead | Incremental error reduction |
|---|---|---|---|---|---|
| L1 | 622 | 604 | 8.2% | 10.1% | 0.5–5 Hz: $58\%$ $[52, 64]$ |
| L2 | 648 | 619 | 7.5% | 9.4% | 5–15 Hz: $43\%$ $[37, 49]$ |
| L3 | 674 | 641 | 6.8% | 8.7% | 15–40 Hz: $31\%$ $[24, 37]$ |
| L4 | 705 | 668 | 6.1% | 8.0% | QRS width: $18\%$ $[11, 24]$ |

Across RVQ levels 1–4, perplexity increases and dead-code rates decrease for both ECG and PPG. The functional probes associate successive residual levels with improvements across progressively higher frequency bands and local QRS morphology. Together with the matched flat-VQ comparison, these results support an actively utilized, coarse-to-fine residual representation without assigning exclusive physiological semantics to individual codebooks.