PPG-LM: A PHOTOPLETHYSMOGRAPHY–LANGUAGE MODEL WITH MULTI-LEVEL CLINICAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Photoplethysmography (PPG) is widely recorded by clinical monitors and consumer wearables, providing a scalable source of continuous physiological information. These recordings offer an opportunity for physiological assessment at scale, but realizing this potential requires models to learn from both signal-derived physiological supervision and broader clinical context captured in electronic health records (EHRs). This involves aligning information spanning local observations, care events, and entire visits with PPG representations at corresponding temporal scales. However, existing PPG foundation models primarily rely on task-specific prediction heads, while the medical knowledge of large language models does not necessarily translate into waveform understanding. To bridge this gap, we introduce **PPG-LM**, the first PPG–language model family to learn physiological representations from both signal-derived supervision and broader clinical context captured in EHRs. To construct clinically grounded captions, we develop an automatic captioning pipeline that generates segment-, event-, and visit-level descriptions from signal measurements and structured EHR records. We then learn from these pairs through a two-stage framework that first establishes segment–language correspondence through contrastive learning and waveform-conditioned captioning, then extends alignment to events and visits through time-aware aggregation and temporal statement matching. Pretrained on approximately 73k hours of PPG, PPG-LM supports language-based recognition, cross-modal retrieval, and segment captioning. Experiments on MC-MED, MIMIC-III, and VitalDB show improved retrieval and caption factuality over language-model baselines and gains over PPG and time-series foundation models on multiple clinical prediction tasks.

1 INTRODUCTION

Photoplethysmography (PPG) is a low-cost optical measure of peripheral blood-volume changes, widely recorded by clinical monitors and wearables (Allen, 2007; Charlton et al., 2022). These recordings provide cardiac and vascular information at scale, motivating foundation models that learn generalizable waveform representations (Pillai et al., 2025; Saha et al., 2025; Nie et al., 2026). However, existing PPG foundation models primarily support task-specific prediction, with limited support for recognizing, retrieving, or describing physiological observations through clinical language. Recent efforts to adapt large language models (LLMs) to PPG question answering (Pham et al., 2026) address a complementary gap: textual medical knowledge alone does not ensure an understanding of pulse morphology and timing (Wang et al., 2026). PPG and Language alignment offers a shared interface for these capabilities, with demonstrated benefits in medical imaging (Huang et al., 2021), wearable sensing (Zhang et al., 2026a), and sleep analysis (Xu et al., 2026). For PPG, electronic health records (EHRs) provide clinical context, including patient presentation, medical history, and care events, that complements signal-derived physiological measurements (He et al., 2026; Kansal et al., 2025). Expressing these complementary sources as language supervision offers an opportunity to learn both waveform characteristics and broader clinical associations.

Training a PPG–language model to learn both fine-grained waveform characteristics and broader clinical information poses two coupled challenges: (i) *Clinically grounded caption generation.* Clinical records contain heterogeneous measurements, histories, and care events but do not directly provide descriptions paired with PPG observations. Constructing useful supervision requires selecting

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.

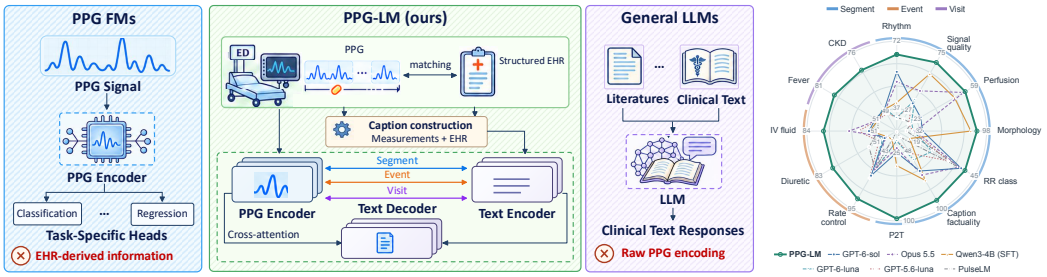


Figure 1: **Motivation and overview of PPG-LM.** **PPG FMs** learn waveform representations with limited integration of language and EHR context. **General LLMs** capture medical knowledge from text but require adaptation to encode raw PPG. **PPG-LM** bridges this gap by pairing PPG with captions derived from physiological measurements and structured EHRs. The radar plot highlights PPG-LM's superior performance over baselines on selected segment-, event-, and visit-level tasks.

and verbalizing information from both signal measurements and EHRs, distinguishing measured waveform properties from documented clinical context, and organizing the resulting descriptions at segment, event, and visit levels. (ii) *Multi-level temporal alignment.* These descriptions correspond to individual windows, observations around care events, and variable-length visits. Event- and visit-level alignment therefore requires aggregating multiple windows, but pooling without temporal information can obscure physiological changes. A model must capture these temporal relationships while maintaining segment–language correspondence as broader clinical supervision is introduced.

To address these challenges, we introduce **PPG-LM**, a family of PPG–language models trained with clinically grounded, multi-level supervision. To construct clinically grounded captions, we develop an automatic captioning pipeline combines signal measurements and structured EHR records to generate segment-, event-, and visit-level descriptions, each paired with PPG observations over the corresponding temporal scope. We then learn these correspondences through a two-stage alignment framework with waveform and text encoders shared across levels. Stage 1 learns segment–language correspondence from captions derived from signal measurements and time-matched vital signs, combining contrastive learning and waveform-conditioned captioning with auxiliary clinical-attribute supervision. Stage 2 extends alignment to medication events and variable-length visits: signed time offsets guide event aggregation, while global and time-anchored visit latents support clinical-caption alignment and temporal statement matching. Our main contributions are as follows:

Table 1: Comparison of PPG–language studies.

Study	Signal		Text Source	Text scope		
	Raw	Context		Segment	Event	Visit
SensorLM (Zhang et al., 2026a)	✗	1 day	Sensor	✓	✗	✗
NormWear (Luo et al., 2024)	✓	6 s	Labels	✓	✗	✗
CSFM (Gu et al., 2026)	✓	10 s	Labels	✓	✗	✗
PulseLM (Pham et al., 2026)	✓	10 s	Labels	✓	✗	✗
CAP (He et al., 2026)	✓	5 min	EHR	✓	✗	✗
PPG-LM	✓	**Variable**	**Signal+EHR**	✓	✓	✓

❶ PPG-LM, a family of PPG–language models. We introduce the first PPG–language model family to learn physiological representations from multi-level clinical language supervision, training on approximately 73,700 hours of PPG to support recognition, retrieval, and captioning.

❷ Clinically grounded multi-level caption generation. We develop an automatic pipeline that combines signal measurements and structured EHR records to generate segment-, event-, and visit-level captions, preserving the evidence source and temporal scope of each description.

❸ Two-stage multi-level alignment. We propose a two-stage framework that first learns segment–language correspondence through contrastive learning and captioning, then extends alignment to event and visit levels through time-aware aggregation and temporal statement matching.

❹ Comprehensive evaluation. Experiments on MC-MED, MIMIC-III, and VitalDB assess recognition, retrieval, captioning, temporal understanding, and clinical prediction against PPG FMs, and LLMs, demonstrating language capabilities and transferable representations across care settings.

2 Related Work

Foundation models for PPG and time series. PPG foundation models learn transferable representations through contrastive and non-contrastive learning with participant and temporal cues (Ab-

baspourazad et al., 2024; Ding et al., 2024), morphology-aware objectives, and uncurated wearable data (Pillai et al., 2025; Saha et al., 2025). Generative approaches use autoregressive prediction and masked reconstruction (Chen et al., 2025; Guo et al., 2026; Zhang et al., 2026b), while cross-signal methods use synchronized ECG and respiratory supervision (Nie et al., 2026; Geenjaar et al., 2026). However, their signal-centered objectives leave waveform–language correspondence unmodeled, and clinical applications typically require task-specific predictors.

LLMs for PPG modeling. LLM-based methods use numerical prompting and parameter-efficient adaptation for physiological estimation (Liu et al., 2024; Ali et al., 2025). PulseLM (Pham et al., 2026) pairs pretrained PPG encoders with instruction-tuned LLMs for question answering, while arrhythmia systems combine signal–text alignment with retrieval-augmented report generation (Huh et al., 2026). Video-based remote PPG methods use text prototypes and physiological prompts for hemodynamic modeling (Xie et al., 2026). These methods center on prescribed estimates, answers, or reports, leaving joint waveform–language learning across clinical temporal levels unaddressed.

Time series and language alignment. Time series–language models pair sensor signals with activity, trend, behavioral, and affective descriptions (Moon et al., 2023; Li et al., 2025; Zhang et al., 2026a; Singh et al., 2024; Chen et al., 2026), and clinical signals with text across ECG, EEG, sleep, and multivariate recordings (Wang et al., 2025; Jiang et al., 2025; Xu et al., 2026; Langer et al., 2026). PPG–EHR alignment uses medical histories as patient-level anchors (He et al., 2026). Multi-level objectives capture local and global semantics in medical imaging and video (Huang et al., 2021; Wang et al., 2022; Ashutosh et al., 2023; Yuan et al., 2024; Bannur et al., 2023), and finer physiological structure in ECG and sleep (Wang et al., 2025; Xu et al., 2026). However, these approaches do not directly support PPG–language alignment across multi-level clinical timescales: none jointly grounds local physiological descriptions, care-event context, and visit-level clinical information.

3 Clinically Grounded Caption Generation

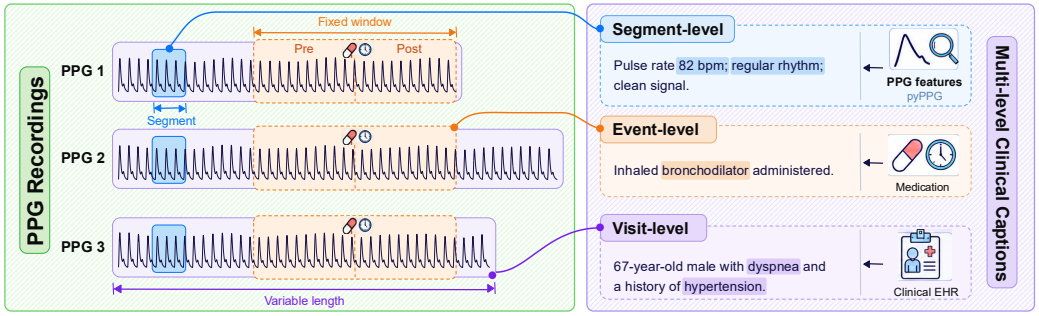


Figure 2: **Clinically Grounded Caption Generation:** (1) *Segment-level* captions describe physiological features and signal quality within 30-second PPG windows. (2) *Event-level* captions provide medication context for fixed windows before and after administration. (3) *Visit-level* captions associate variable-length recordings with demographics, and medical history from clinical EHRs.

To train PPG-LM, we construct a corpus from MC-MED (Kansal et al., 2025), which pairs PPG recordings with structured electronic health records (EHRs) from emergency-department visits. We generate segment-, event-, and visit-level captions from signal measurements and EHRs, pairing them with PPG observations at the corresponding temporal scales (Figure 2). Captions use fixed language templates with reproducible template selection and sentence shuffling. Corpus details, captioning rules, and template examples are provided in Appendices D.1, B, and E, respectively.

***Segment-level* captions.** Using pyPPG (Goda et al., 2024), we describe up to eight attributes per PPG window: pulse rate, rate trend, rhythm regularity, pulsatile amplitude, signal quality, augmentation-related morphology, a vascular-response index, and beat-to-beat amplitude variability. Available fields yield numerical or qualitative descriptions.

***Event-level* captions.** We extract administration timestamps, drug names, and routes from EHR medication records. Route filtering and drug-name rules map eligible records to ten classes; near-duplicates are removed within each visit and class. For an administration at time a, we select up to three latest windows in $[a-30, a)$ and up to five evenly spaced windows in $[a, a+60]$ (times

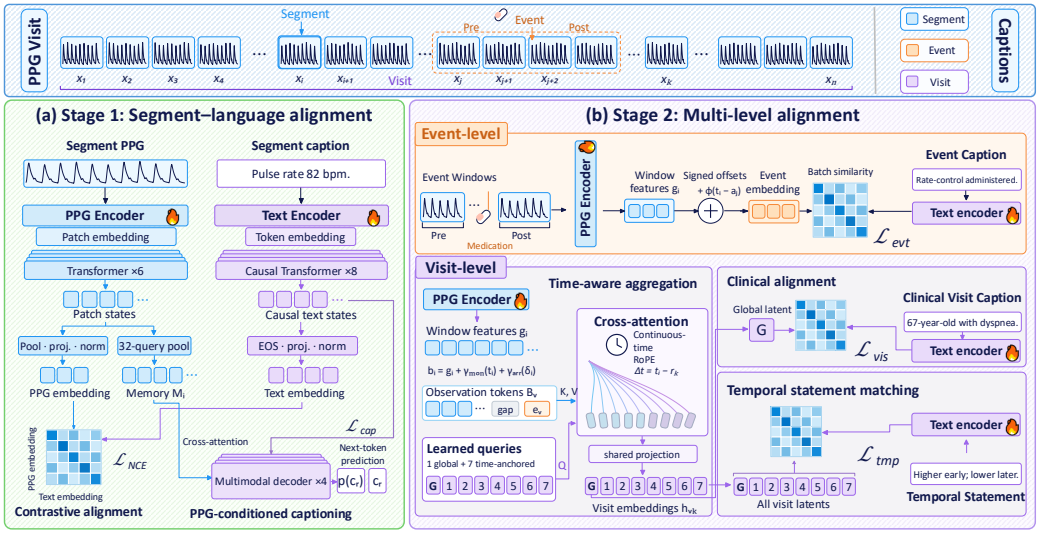


Figure 3: **Two-stage multilevel alignment. (a)** Stage 1 learns PPG and text encoders through contrastive segment–language alignment and PPG-conditioned caption generation. **(b)** Stage 2 extends the shared encoders to event- and visit-level alignment.

in minutes). The caption names the administered class and is paired with these windows. Controls follow the same sampling rule within the visit, excluding anchors near recorded administrations.

Visit-level **captions.** We combine encounter records and demographics to summarize age, sex, arrival mode, triage acuity, normalized chief complaints, and first recorded vital signs. Prior-diagnosis codes are mapped to clinical descriptions, deduplicated, and ranked by relevance and recency; active home medications are summarized by class. These fields form one caption paired with observations across the encounter. Ordered segment rates and rhythm labels additionally yield temporal statements of state *occurrence*, *proportion*, contiguous *duration*, and rate *ordering*.

4 METHOD

The captions in Section 3 pair local physiological measurements and broader clinical context with PPG observations at different temporal scopes. Learning from these pairs requires both a window-level language interface and representations that combine observations across time. PPG-LM therefore uses shared waveform and text encoders with scope-specific read-outs, trained in two stages (Figure 3). Stage 1 establishes segment–language correspondence through contrastive learning and captioning. Stage 2 learns to aggregate window features for event and visit descriptions, adding temporal statement supervision while retaining the segment objective.

Problem definition. A visit is a time-ordered sequence $V = ((x_i, t_i))_{i=1}^{n_V}$, where $x_i \in \mathbb{R}^{3750}$ is a standardized 30-second PPG window and t_i is its start time in minutes since the first window. A caption c^ℓ, $\ell \in \{\text{seg}, \text{evt}, \text{vis}\}$, is paired with a single window, windows around a medication or control anchor, or observations distributed across a visit, respectively. In addition to its clinical caption, a visit may have temporal statements $T_V = \{d_{Vr}\}_{r=1}^{m_V}$ describing state occurrence, proportion, contiguous duration, and rate ordering.

PPG and text encoders. We share both encoders across levels to reuse waveform and language features, while adapting their read-outs to each description's scope. The PPG encoder f_θ divides a window into non-overlapping patches, adds learned positional embeddings, and applies Transformer blocks (Vaswani et al., 2017). Attention pooling summarizes the resulting patch states H_i into a window feature g_i. The text encoder ψ_ϕ is a causal Transformer using CLIP's byte-pair encoding (BPE) vocabulary (Radford et al., 2021). A learned prefix e_ℓ identifies the description level, and a level-specific projection P_ℓ maps the end-of-sequence (EOS) state to a normalized text embedding: $g_i = f_\theta(x_i)$, $u^\ell(c) = \text{norm}(P_\ell\, \psi_\phi([e_\ell; c])_{\text{EOS}})$, where $\text{norm}(v) = v/\|v\|_2$. Waveform read-outs operate on g_i for segment alignment and on collections of window features for event and visit alignment. Encoder details are given in Appendix C.1.

4.1 Stage 1: Segment–Language Pretraining

Following CoCa (Yu et al., 2022), Stage 1 combines contrastive alignment of windows and descriptions with token-level caption supervision. For a batch of B window–caption pairs (x_i, c_i) from Section 3, we form normalized waveform embeddings $z_i^{\text{seg}} = \text{norm}(W_{\text{seg}} g_i)$ and logits $Q_{ij} = s_{\text{seg}}(z_i^{\text{seg}})^\top u^{\text{seg}}(c_j)$, with learned scale $s_{\text{seg}} > 0$. To avoid treating captions with identical binned measurements as negatives, we assign content keys κ_i and use $\mathcal{C}_i^{\text{seg}} = \{i\} \cup \{j : \kappa_j \neq \kappa_i\}$ within the batch. For candidate sets $\mathcal{C}$, InfoNCE (van den Oord et al., 2018) in CLIP's symmetric form (Radford et al., 2021) is

$$\mathcal{L}_{\text{NCE}}(Q; \mathcal{C}) = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(Q_{ii})}{\sum_{j \in \mathcal{C}_i} \exp(Q_{ij})} + \log \frac{\exp(Q_{ii})}{\sum_{j \in \mathcal{C}_i} \exp(Q_{ji})} \right]. \tag{1}$$

The two terms align signals and text in both directions; later levels reuse this loss with their respective logits and candidate sets.

For complementary token-level supervision, a separate attention pooler forms waveform memory M_i from the shared patch states H_i. A multimodal decoder receives causal token states from the shared text encoder and cross-attends to M_i to predict caption tokens. With teacher forcing, we minimize $\mathcal{L}_{\text{cap}} = -N_{\text{tok}}^{-1} \sum_{i,r} \log p_{\phi,\eta}(c_{i,r} \mid c_{i,<r}, M_i)$ over the N_{tok} non-padding targets, including EOS; η denotes the additional captioning parameters. The combined objective $\mathcal{L}_{\text{seg}} = \mathcal{L}_{\text{NCE}}(Q; \mathcal{C}^{\text{seg}}) + \mathcal{L}_{\text{cap}}$ trains both shared encoders from scratch and is retained in Stage 2.

4.2 Stage 2: Multi-Level Alignment

Event and visit descriptions concern multiple observations. Stage 2 therefore adds temporal aggregation over the shared window features and aligns the resulting representations with descriptions at the corresponding scope.

Event-level alignment. Event captions pair a medication class or no-administration control with nearby PPG windows (Section 3). For view j, we preserve timing relative to its anchor a_j by adding signed offset embeddings $\varphi(t_i - a_j)$ and pooling features over selected window indices E_j with a learned query:

$$z_j^{\text{evt}} = \text{norm}\big(W_{\text{evt}} \,\text{Attn}\left(q_{\text{evt}}; \{g_i + \varphi(t_i - a_j)\}_{i \in E_j}\right)\big). \tag{2}$$

Offsets are measured in minutes on the visit clock, distinguishing observations before and after the anchor and their temporal distance. Using Equation 1, we minimize $\mathcal{L}_{\text{evt}} = \mathcal{L}_{\text{NCE}}(Q^{\text{evt}}; \mathcal{C}^{\text{evt}})$, with logits $Q_{jk}^{\text{evt}} = s_{\text{evt}}(z_j^{\text{evt}})^\top u^{\text{evt}}(c_k^{\text{evt}})$ and learned scale $s_{\text{evt}} > 0$. Candidate sets retain the matched pair and exclude other views of the same class to avoid contrasting identical treatment categories; controls form one class. Attention details appear in Appendix C.3.

Visit-level alignment. A clinical visit caption summarizes the encounter, while temporal statements may require evidence from different parts of its trajectory. To bound computation, we select at most 256 windows indexed by $\mathcal{I}_V$ across the monitoring span. We encode monitoring time t_i and time since ED arrival δ_i as $b_i = g_i + \gamma_{\text{mon}}(t_i) + \gamma_{\text{arr}}(\delta_i)$. The input $B_V = [(b_i)_{i \in \mathcal{I}_V}; G_V; e_V]$ also includes gap tokens G_V for monitoring interruptions and an evidence token e_V for coverage and quality.

Two attention blocks aggregate B_V through learned queries (Jaegle et al., 2021). We fix their number at eight: one global query and seven with reference times r_k relative to the first window (Shukla & Marlin, 2021). We apply RoFormer's rotary position embeddings (RoPE) (Su et al., 2024) to actual time offsets in minutes. For an anchored query q_k and a window key k_i,

$$\tilde{q}_k^\top \tilde{k}_i = (q_k^{\text{rot}})^\top R_\Theta(t_i - r_k) k_i^{\text{rot}} + (q_k^{\text{free}})^\top k_i^{\text{free}}, \qquad k = 1, \ldots, 7. \tag{3}$$

Here R_Θ rotates the rot channels and free channels remain unchanged. All queries attend to every valid input token; anchors do not restrict attention to fixed intervals. The output latents ℓ_{Vk} share a projection $h_{Vk} = \text{norm}(W_{\text{lat}} \ell_{Vk})$, $k = 0, \ldots, 7$. We align the global embedding $z_V^{\text{vis}} = h_{V0}$ with $u^{\text{vis}}(c_V^{\text{vis}})$ through in-batch InfoNCE, $\mathcal{L}_{\text{vis}}$, and use all eight h_{Vk} for temporal statement matching. Encoding and anchor details appear in Appendix C.3.

Temporal statement matching. We adapt ColBERT's late interaction (Khattab & Zaharia, 2020) to match each statement to its most compatible visit latent. For trajectory i and a visit j with

$m_j > 0$ available statements, let $v_{jr} = u^{\text{vis}}(d_{jr})$: $R_{ij} = \frac{1}{m_j} \sum_{r=1}^{m_j} \max_{k=0,\dots,7} h_{ik}^{\top} v_{jr}$. The mean normalizes for statement count, and each statement can select a different latent without a prescribed statement-to-anchor assignment. Alongside in-batch alignment, we rank the true set above a version i^- with one incompatible statement:

$$\mathcal{L}_{\text{tmp}} = \mathcal{L}_{\text{NCE}}(s_{\text{vis}} R; \mathcal{C}^{\text{full}}) + \tfrac{1}{2}\, \mathbb{E}_i \big[-\log \sigma\big(s_{\text{vis}}(R_{ii} - R_{i,i^-})\big) \big], \tag{4}$$

where σ is the logistic function and $\mathcal{C}^{\text{full}}$ admits all in-batch pairs. Changing one statement preserves most of the description while making its temporal content the discriminating target. Missing statements and unavailable corruptions are masked.

Stage 2 initializes from Stage 1 and alternates segment, visit, and event batches, retaining segment supervision to preserve local alignment (Ashutosh et al., 2023). For batch type $b \in \{\text{seg}, \text{evt}, \text{vis}\}$ and windows $\mathcal{Q}$ re-encoded with gradients, the alignment objective is

$$\mathcal{L}^{(b)} = \mathcal{L}_{\text{seg}}(\mathcal{Q}) + \lambda_{\text{dst}} \mathcal{L}_{\text{dst}}(\mathcal{Q}) + \mathbb{I}_{b=\text{evt}} \mathcal{L}_{\text{evt}} + \mathbb{I}_{b=\text{vis}} (\mathcal{L}_{\text{vis}} + \mathcal{L}_{\text{tmp}}). \tag{5}$$

Here $\mathbb{I}$ is an indicator, and $\mathcal{L}_{\text{dst}}$ is the mean cosine distance to frozen Stage-1 window embeddings. Training schedules, hyperparameter settings, and inference procedures are given in Appendix C.5.

5 Experiments

Datasets. We use MC-MED (Kansal et al., 2025) for pretraining and in-domain evaluation, with patient-disjoint training, validation, and test splits. MC-MED pairs PPG recordings from Stanford's adult emergency department (ED) with structured EHRs. For external evaluation, we use MIMIC-III (Johnson et al., 2016), which links intensive-care bedside waveforms from Beth Israel Deaconess Medical Center to clinical records, and VitalDB (Lee et al., 2022), which pairs intraoperative biosignals from Seoul National University Hospital with perioperative clinical information.

Baselines. For language-based tasks, we compare with four general-purpose LLMs—GPT-5.6-luna (OpenAI, 2026a), GPT-6-luna (OpenAI, 2026b), GPT-6-sol (OpenAI, 2026c), and Claude Opus 5.5 (Anthropic, 2026). We also include the Qwen2.5-7B-Instruct variant of PulseLM (Pham et al., 2026), trained with a PaPaGei PPG encoder (Pillai et al., 2025). Separately, we train a Qwen3-4B (SFT) baseline (Qwen Team, 2025) using a frozen PaPaGei-S encoder and supervised fine-tuning on PPG signals paired with captions generated by our captioning pipeline. For clinical prediction, we include PPG foundation models PaPaGei-S (Pillai et al., 2025), Pulse-PPG (Saha et al., 2025), AnyPPG (Nie et al., 2026), and SIGMA-PPG (Guo et al., 2026), together with general time-series foundation models MOMENT-L (Goswami et al., 2024) and Chronos-2 (Ansari et al., 2025).

5.1 Zero-Shot Physiological Understanding

Table 2: **Zero-shot physiological recognition.** Macro-F1 (↑) for rhythm, quality (Qual.), perfusion (Perf.), morphology (Morph.), and respiratory-rate class (RR). Bold marks the best in each column.

Model	MC-MED					MIMIC-III			VitalDB		
	Rhythm↑	Qual.↑	Perf.↑	Morph.↑	RR↑	Rhythm↑	Qual.↑	RR↑	Rhythm↑	Qual.↑	RR↑
GPT-5.6-luna	0.322	0.314	0.174	0.259	0.332	0.376	0.291	0.266	0.317	0.162	0.313
GPT-6-luna	0.319	0.361	0.193	0.263	0.259	0.357	0.315	0.253	0.326	0.175	0.307
GPT-6-sol	0.560	0.332	0.250	0.354	0.390	0.572	0.295	0.380	0.375	0.128	0.326
Opus 5.5	0.508	0.436	0.521	0.284	0.380	0.598	0.343	0.397	0.330	0.228	**0.330**
PulseLM	0.319	0.199	0.188	0.195	0.405	0.242	0.297	0.354	0.181	0.326	0.165
Qwen3-4B (SFT)	0.399	0.578	0.464	0.826	0.158	0.272	0.564	0.171	0.231	0.515	0.015
PPG-LM	**0.651**	**0.673**	**0.531**	**0.886**	**0.407**	**0.670**	**0.786**	**0.428**	**0.389**	**0.583**	0.329

Zero-shot physiological recognition. We test recognition of physiological categories from PPG in Table 2. Macro-F1 is reported for rhythm regularity, signal quality (Qual.), waveform-derived perfusion strength (Perf.) and augmentation-related morphology (Morph.), and respiratory-rate class (RR). PPG-LM achieves the highest scores on 10 of 11 dataset–task pairs, including rhythm and quality on all three datasets and both perfusion and morphology on MC-MED. The exception is VitalDB RR, where it closely matches Opus 5.5 (0.329 vs. 0.330). These results support recognition of multiple waveform attributes through a shared language interface, with transfer across emergency, intensive-care, and surgical settings.

Zero-shot physiological estimation. We assess whether generated descriptions preserve physiological facts and numerical measurements in Table 3. Fact measures the accuracy of stated facts across six shared attributes; HR, RR, SBP, and MAP report MAE for heart rate, respiratory rate, systolic blood pressure, and mean arterial pressure, respectively. PPG-LM achieves the highest fact accuracy on MC-MED, MIMIC-III, and VitalDB (0.923, 0.920, and 0.820) and the lowest reported errors for all displayed measurements. HR MAE is 3.90–4.27 bpm against an independent ECG reference, supporting physiological agreement beyond the PPG-derived training captions. RR errors are computed on valid outputs, so differences in output coverage limit their direct comparison.

Table 3: **Zero-shot physiological estimation.** Fact is the accuracy of stated facts across six shared attributes (↑). Numerical columns report MAE (↓): HR in bpm, RR in breaths/min, and SBP/MAP in mmHg. MAE uses valid numerical outputs; bold marks the best reported value in each column.

Model	MC-MED					MIMIC-III			VitalDB		
	HR↓	RR↓	SBP↓	MAP↓	Fact↑	HR↓	RR↓	Fact↑	HR↓	RR↓	Fact↑
GPT-5.6-luna	10.06	3.95	21.73	15.12	0.499	9.42	5.15	0.456	8.49	2.99	0.486
GPT-6-luna	11.24	3.93	22.24	14.51	0.559	11.52	4.71	0.533	8.86	2.34	0.579
GPT-6-sol	4.58	3.85	21.56	14.48	0.659	5.21	5.10	0.660	5.51	2.26	0.619
Opus 5.5	4.67	3.32	19.80	14.26	0.696	4.37	3.99	0.647	4.88	2.44	0.642
PulseLM	29.20	4.21	20.44	15.47	0.356	29.44	4.59	0.440	37.12	7.03	0.481
Qwen3-4B (SFT)	7.36	5.10	20.14	13.93	0.743	8.54	5.20	0.665	8.98	7.80	0.674
PPG-LM	**3.90**	**3.10**	**12.50**	**6.17**	**0.923**	**4.27**	**3.83**	**0.920**	**3.98**	**2.23**	**0.820**

Zero-shot cross-modal retrieval. We test whether a PPG signal retrieves its paired description (signal-to-text, S2T) and vice versa (text-to-signal, T2S). Table 4 reports R@1, the fraction of queries whose paired item ranks first, averaged over two 100-candidate pools. Following SensorLM and SleepLM (Zhang et al., 2026a; Xu et al., 2026), we evaluate general-purpose LLMs only in S2T to avoid the long numerical contexts required to include all 100 candidate waveforms in a T2S prompt. PPG-LM leads both directions across all three datasets, with S2T R@1 of 0.975, 0.805, and 0.555 on MC-MED, MIMIC-III, and VitalDB, respectively; the corresponding T2S scores are 0.965, 0.755, and 0.550. These results support bidirectional signal–text alignment that transfers to external datasets, with lower retrieval accuracy outside MC-MED.

Table 4: **Zero-shot cross-modal retrieval.** R@1 with 100 candidates. Bold: best.

Model	MC-MED		MIMIC-III		VitalDB500	
	S2T↑	T2S↑	S2T↑	T2S↑	S2T↑	T2S↑
GPT-5.6-luna	0.105	—	0.090	—	0.075	—
GPT-6-luna	0.110	—	0.085	—	0.085	—
GPT-6-sol	0.240	—	0.260	—	0.200	—
Opus 5.5	0.190	—	0.275	—	0.230	—
Qwen3-4B (SFT)	0.385	0.600	0.160	0.185	0.125	0.180
PPG-LM	**0.975**	**0.965**	**0.805**	**0.755**	**0.555**	**0.550**

5.2 Zero-Shot Event- and Visit-Level Physiological Recognition

Table 5: **Zero-shot event-level physiological recognition.** Class-versus-control AUROC (↑) for the five highest-scoring MC-MED classes for PPG-LM and all available VitalDB classes. Bold marks the best score in each column.

Model	MC-MED					VitalDB			
	Rate ctrl.↑	Vasopress.↑	Vasodil.↑	IV fluid↑	Diuretic↑	Opioid↑	Sedative↑	Vasopress.↑	Vasodil.↑
GPT-5.6-luna	0.674	0.707	0.553	0.587	0.499	0.567	0.641	0.578	0.614
GPT-6-luna	0.626	0.715	0.648	0.530	0.478	0.600	0.594	0.480	0.693
GPT-6-sol	0.651	0.699	0.648	0.536	0.467	0.663	0.679	0.480	0.585
Opus 5.5	0.595	0.707	0.636	0.633	0.460	0.722	0.766	0.571	0.634
PulseLM	0.591	0.589	0.458	0.493	0.553	0.591	0.613	0.602	0.480
Qwen3-4B (SFT)	0.324	0.464	0.507	0.531	0.453	0.532	0.386	0.503	0.481
PPG-LM	**0.860**	**0.830**	**0.772**	**0.759**	**0.754**	**0.731**	**0.807**	**0.612**	**0.741**

Zero-shot event-level physiological recognition. We distinguish observations around medication administrations from no-administration controls (Table 5). Each column reports class-versus-control AUROC for a medication class: rate-control agents (Rate ctrl.), vasopressors (Vasopress.), vasodilators (Vasodil.), intravenous fluids (IV fluid), diuretics, opioids, or sedatives. PPG-LM outperforms the reported baselines on all displayed classes, with AUROC of 0.754–0.860 for the five selected

MC-MED classes and 0.612–0.807 for the four VitalDB classes. These results support recognition of documented treatment context across emergency and surgical settings.

Table 6: **Zero-shot visit-level recognition.** Balanced accuracy (↑) for home-medication use (MC-MED) and clinical status (VitalDB). PPG-LM uses eight-window fact embeddings with held-out threshold calibration. Bold marks the best score in each column.

Model	MC-MED				VitalDB			
	CCB↑	AntiHTN↑	AntiPL↑	AntiDM↑	HTN↑	DM↑	Anemia↑	eGFR↑
GPT-5.6-luna	0.468	0.509	0.477	0.520	0.500	0.465	0.487	0.500
GPT-6-luna	0.497	0.484	0.421	0.464	0.475	0.507	0.495	0.498
GPT-6-sol	0.500	0.500	0.505	0.507	**0.523**	0.502	0.523	0.500
Opus 5.5	0.500	0.443	0.468	0.507	0.519	0.512	0.484	0.500
PulseLM	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Qwen3-4B (SFT)	0.535	0.540	0.438	0.444	0.500	0.500	0.500	0.500
PPG-LM	**0.558**	**0.570**	**0.544**	**0.532**	0.520	**0.562**	**0.531**	**0.559**

Zero-shot visit-level recognition. We assess whether eight PPG windows identify recorded home-medication use and clinical status, reporting balanced accuracy in Table 6. The MC-MED columns denote calcium-channel blockers (CCB), antihypertensives (AntiHTN), antiplatelets (AntiPL), and antidiabetic drugs (AntiDM). VitalDB columns cover hypertension (HTN), diabetes (DM), anemia, and reduced kidney function (eGFR below 60 mL/min/1.73 m^2). With held-out threshold calibration, PPG-LM leads on seven of eight tasks; the exception is HTN (0.520 vs. 0.523 for GPT-6-sol). Balanced accuracy of 0.520–0.570 indicates modest discrimination of these clinical facts. Additional MC-MED evaluations cover eight triage attributes and eleven medical-history conditions in Appendix D.5, Table 21(a) and (b), respectively.

5.3 Clinical Prediction with Frozen Representations

Table 7: **Clinical prediction with linear probes.** Three classification and three regression tasks per dataset (mean ± SD). Bold: best mean; †: window-level target.

Task	Foundation models						PPG-LM	
	PaPaGei-S	PulsePPG	AnyPPG	SIGMA-PPG	MOMENT-L	Chronos-2	Stage 1	Stage 2
MC-MED (ED) — AUROC ↑								
Sex	0.692±0.006	0.719±0.007	0.783±0.008	0.712±0.008	0.724±0.005	0.739±0.008	0.813±0.006	**0.856**±0.005
ED admission	0.673±0.011	0.681±0.009	0.750±0.013	0.679±0.011	0.700±0.011	0.728±0.013	0.765±0.012	**0.787**±0.010
Atrial fibrillation	0.742±0.012	0.752±0.013	0.845±0.010	0.755±0.013	0.768±0.014	0.815±0.012	0.857±0.010	**0.870**±0.011
MC-MED (ED) — R^2 ↑								
Age	0.395±0.016	0.428±0.018	0.626±0.020	0.389±0.017	0.475±0.017	0.521±0.016	0.691±0.016	**0.782**±0.014
Hemoglobin	0.083±0.011	0.098±0.013	0.196±0.013	0.102±0.012	0.123±0.014	0.164±0.014	0.238±0.014	**0.274**±0.012
Glucose	0.032±0.007	0.039±0.007	0.065±0.008	0.041±0.009	0.028±0.007	0.049±0.011	0.071±0.007	**0.083**±0.011
MIMIC-III (ICU) — AUROC ↑								
Sex	0.651±0.024	0.686±0.014	0.713±0.018	0.672±0.016	0.682±0.019	0.664±0.020	0.730±0.019	**0.733**±0.018
In-hospital mortality	0.697±0.029	0.741±0.034	0.756±0.031	0.732±0.025	0.751±0.025	0.772±0.020	0.786±0.022	**0.793**±0.021
Circulatory diagnosis	0.666±0.015	0.690±0.022	0.725±0.017	0.658±0.018	0.672±0.012	0.677±0.017	0.720±0.022	**0.726**±0.023
MIMIC-III (ICU) — R^2 ↑								
Age	0.295±0.039	0.308±0.031	0.468±0.029	0.288±0.036	0.347±0.036	0.367±0.028	0.502±0.030	**0.507**±0.028
BUN	0.029±0.011	0.029±0.009	0.065±0.019	0.034±0.015	0.055±0.015	0.074±0.012	0.081±0.012	**0.086**±0.013
Hemoglobin	0.035±0.010	0.048±0.010	0.079±0.023	0.034±0.009	0.046±0.012	0.076±0.018	0.092±0.024	**0.095**±0.020
VitalDB (OR) — AUROC ↑								
ASA ≥ 3	0.606±0.021	0.642±0.020	0.690±0.026	0.605±0.035	0.625±0.035	0.655±0.040	0.699±0.034	**0.704**±0.025
Hypertension	0.598±0.016	0.607±0.020	0.635±0.019	0.595±0.025	0.610±0.016	0.617±0.023	**0.656**±0.016	0.655±0.017
ICU admission	0.690±0.009	0.700±0.004	0.747±0.002	0.683±0.009	0.662±0.011	0.714±0.009	0.733±0.004	**0.751**±0.006
VitalDB (OR) — R^2 ↑								
Age	0.240±0.026	0.263±0.036	0.403±0.024	0.259±0.026	0.302±0.025	0.337±0.032	0.418±0.030	**0.435**±0.029
$EtCO_2$†	0.098±0.023	0.121±0.034	0.105±0.028	0.113±0.024	0.133±0.029	0.150±0.031	**0.153**±0.038	0.152±0.037
BIS†	0.380±0.019	0.398±0.022	0.393±0.021	0.387±0.020	0.402±0.021	0.404±0.024	0.407±0.022	**0.408**±0.022

We evaluate frozen PPG representations on 32 dataset–task pairs across MC-MED, MIMIC-III, and VitalDB using logistic and ridge regression with patient-disjoint splits. All models use the same four windows per visit: baselines and Stage 1 mean-pool features, while Stage 2 uses its learned aggregator. Stage 2 PPG-LM exceeds all six foundation-model baselines in mean performance on all 32 tasks and Stage 1 on 30, except VitalDB hypertension and $EtCO_2$, although glucose prediction remains weak ($R^2 = 0.083$). Table 7 presents selected results; Appendix D.7, Table 22 provides complete results and protocols.

5.4 Model Analysis

Ablation studies. Table 8 shows that temporal matching and event offsets support temporal and event discrimination, respectively. Continued segment supervision preserves retrieval, while distillation helps mainly when it is removed. Comparing CLIP and SigLIP, CoCa gives the best Stage-2 rhythm, quality recognition and retrieval (Appendix D.9).

Table 8: **Ablation studies.**

Variant	Temporal acc.↑	Event AUROC↑	Segment R@1↑
w/o temporal loss	0.511	0.698	0.446
w/o signed offsets	0.923	0.617	0.445
w/o segment loss	0.922	0.697	0.202
w/o distillation	0.920	0.700	0.445
w/o seg. loss + distill.	0.919	0.699	0.015
PPG-LM	**0.925**	**0.707**	**0.447**

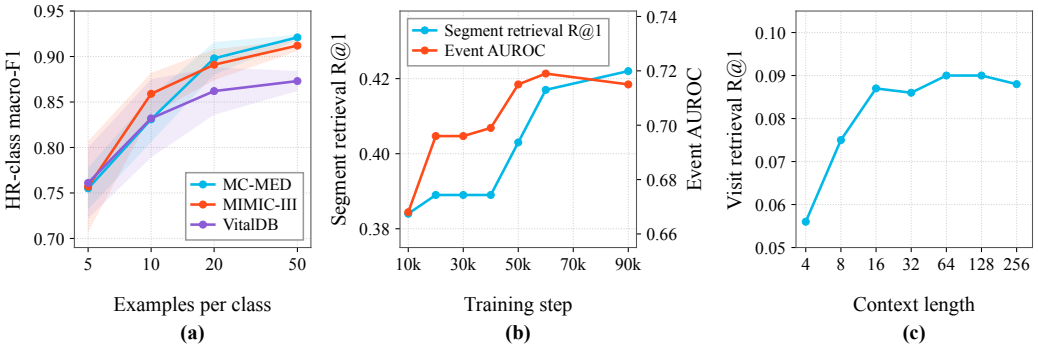


Figure 4: **Model analysis.** (a) HR-class macro-F1 with labeled examples per class; shading denotes ± 1 SD over five support draws. VitalDB uses the 40 Hz waveform view. (b) Segment retrieval and medication-event AUROC during training. (c) Visit retrieval versus context length in windows.

Few-shot adaptation. Figure 4(a) shows HR-class prediction with linear probes on frozen PPG-LM representations. Mean macro-F1 increases consistently as labeled examples per class grow from 5 to 50 on MC-MED, MIMIC-III, and VitalDB. Appendix D.8 reports complete results across tasks and label budgets, along with the separate LLM in-context protocol, whose label budget and adaptation procedure differ from those of the probes.

Training dynamics. Figure 4(b) tracks an extended Stage-2 CoCa run. Segment retrieval improves with training, while medication-event AUROC largely plateaus by 50–60k steps, indicating different benefits from additional optimization across the two tasks.

Visit context. With the Stage-2 CoCa model fixed, increasing context from four to sixteen windows improves visit retrieval, with little further gain from longer contexts (Figure 4(c)). This suggests diminishing returns from additional observations in this retrieval setting.

6 Conclusion

We presented PPG-LM, a PPG–language model family that connects waveforms with clinical context. Our framework combines signal- and EHR-derived captions with two-stage alignment across segment, event, and visit levels. Experiments on MC-MED, MIMIC-III, and VitalDB support its effectiveness in language-based recognition, retrieval, captioning, and clinical prediction.

AI USE STATEMENT

During manuscript preparation, we used generative AI tools solely for language polishing, including grammar correction, sentence refinement, and improvements to clarity and readability. The authors reviewed all suggested edits to ensure that the revised text preserved the intended technical meaning and accurately reflected the methods and results. The authors retained final editorial control and take full responsibility for the manuscript's content, claims, and conclusions.

ETHICS STATEMENT

This study involves secondary analysis of de-identified data from MC-MED, MIMIC-III, and VitalDB (Kansal et al., 2025; Johnson et al., 2016; Lee et al., 2022), in accordance with their respective access requirements and data-use agreements, with no new patient recruitment. PPG-LM is intended for research and has not been validated for clinical decision-making. Medication-event predictions reflect associations and do not establish causal treatment effects. Prospective validation and assessment of subgroup performance and generalizability are needed before clinical deployment, as discussed in Appendix A.

REPRODUCIBILITY STATEMENT

Sections 3–5 describe caption construction, the model framework, and the experiments. Appendices B and E provide captioning rules, language templates, and evaluation prompts, while Appendix C details the architecture, training objectives, hyperparameters, and computational setup. Dataset preprocessing, patient-disjoint splits, baseline configurations, task definitions, and evaluation protocols are documented in Appendix D.

REFERENCES

Salar Abbaspourazad, Oussama Elachqar, Andrew C. Miller, Saba Emrani, Udhyakumar Nallasamy, and Ian Shapiro. Large-scale training of foundation models for wearable biosignals. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pC3WJHf51j.

Syed Anas Ali, Muhammad Wasim Nawaz, Junaid Rashid, Ali Hamid Mahmood, Jungeun Kim, Muhammad Mahboob Ur Rahman, and Qammer H. Abbasi. PEFT QLORA-based fine-tuning of foundation models for vitals estimation using PPG and ECG-based medical IoT data: A feasibility study. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 468–477. IEEE Computer Society, 2025. doi: 10.1109/ICDMW69685.2025.00059.

John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28(3):R1–R39, 2007. doi: 10.1088/0967-3334/28/3/R01.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, Mononito Goswami, Shubham Kapoor, Danielle C. Maddix, Pablo Guerron, Tony Hu, Junming Yin, Nick Erickson, Prateek Mutalik Desai, Hao Wang, Huzefa Rangwala, George Karypis, Yuyang Wang, and Michael Bohlke-Schneider. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025. doi: 10.48550/arXiv.2510.15821.

Anthropic. Introducing Claude Opus 5.5. https://www.anthropic.com/claude-opus-5-5, September 2026. Released September 22, 2026. Accessed September 25, 2026.

Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. HierVL: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23066–23078, 2023.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay.

Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15016–15027, 2023.

Peter H. Charlton, Panicos A. Kyriacou, Jonathan Mant, Vaidotas Marozas, Phil Chowienczyk, and Jordi Alastruey. Wearable photoplethysmography for cardiovascular monitoring. *Proceedings of the IEEE*, 110(3):355–381, 2022. doi: 10.1109/JPROC.2022.3149785.

Yuliang Chen, Arvind Pillai, Yu Yvonne Wu, Tess Z. Griffin, Lisa Marsch, Michael V. Heinz, Nicholas C. Jacobson, and Andrew Campbell. Learning transferable sensor models via language-informed pretraining. *arXiv preprint arXiv:2603.11950*, 2026. doi: 10.48550/arXiv.2603.11950. URL https://arxiv.org/abs/2603.11950.

Zhaoliang Chen, Cheng Ding, Saurabh Kataria, Runze Yan, Minxiao Wang, Randall J. Lee, and Xiao Hu. GPT-PPG: A GPT-based foundation model for photoplethysmography signals. *Physiological Measurement*, 46(5):055004, 2025. doi: 10.1088/1361-6579/add988.

Cheng Ding, Zhicheng Guo, Zhaoliang Chen, Randall J. Lee, Cynthia Rudin, and Xiao Hu. SiamQuality: A ConvNet-based foundation model for photoplethysmography signals. *Physiological Measurement*, 45(8):085004, 2024. doi: 10.1088/1361-6579/ad6747.

Eloy Geenjaar, Vince D. Calhoun, Scott Daly, Gouthaman KV, Lie Lu, Trisha Mittal, and Daniel P. Darcy. A robust PPG foundation model using multimodal physiological supervision. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64846.

Márton Á. Goda, Peter H. Charlton, and Joachim A. Behar. pyPPG: A Python toolbox for comprehensive photoplethysmography signal analysis. *Physiological Measurement*, 45(4):045001, 2024. doi: 10.1088/1361-6579/ad33a2.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16115–16152. PMLR, 2024.

Xiao Gu, Wei Tang, Jinpei Han, Veer Sangha, Fenglin Liu, Shreyank N. Gowda, Antonio H. Ribeiro, Patrick Schwab, Kim Branson, Lei Clifton, Antonio Luiz P. Ribeiro, Zhangdaihong Liu, and David A. Clifton. Cardiac health assessment across scenarios and devices using a multimodal foundation model pretrained on data from 1.7 million individuals. *Nature Machine Intelligence*, 8 (2):220–233, 2026. doi: 10.1038/s42256-026-01180-5. URL https://www.nature.com/articles/s42256-026-01180-5.

Zongheng Guo, Tao Chen, Yang Jiao, Yi Pan, Xiao Hu, and Manuela Ferrario. SIGMA-PPG: Statistical-prior informed generative masking architecture for PPG foundation model. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64268.

Chenyang He, Xinyi Shao, Shun Huang, Bosong Huang, Daoqiang Zhang, Ming Jin, and Cheng Ding. CAP: Towards PPG universal representation learning with patient-level supervision. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 11050–11060. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3818881.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021.

Youngho Huh, Minhwan Noh, Dongwoo Ji, Yuna Oh, and Sukkyu Sun. Multimodal PPG-based arrhythmia detection using a CLIP-initialized multi-task U-Net and LLM-assisted reporting. *Sensors*, 26(8):2316, 2026. doi: 10.3390/s26082316.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021.

Wei-Bang Jiang, Yansen Wang, Bao-Liang Lu, and Dongsheng Li. Neurolm: A universal multi-task foundation model for bridging the gap between language and eeg signals. In *International Conference on Learning Representations*, volume 2025, pp. 55436–55457, 2025.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. doi: 10.1038/sdata.2016.35.

Aman Kansal, Emma Chen, Boyang Tom Jin, Pranav Rajpurkar, and David A. Kim. MC-MED, multimodal clinical monitoring in the emergency department. *Scientific Data*, 12(1):1094, 2025. doi: 10.1038/s41597-025-05419-5.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48. Association for Computing Machinery, 2020. doi: 10.1145/3397271.3401075.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A. Xu, Winnie Chow, Martin Maritsch, Robert Jakob, Ning Wang, Juncheng Liu, Aradhana Verma, Brian Han, Daniel Kim, Henry Chubb, Scott Ceresnak, Aydin Zahedivash, Alexander Sandhu, Fatima Rodriguez, Daniel McDuff, Elgar Fleisch, Oliver Aalami, Filipe Barata, and Paul Schmiedmayer. OpenTSLM: Time-series language models for reasoning over multivariate medical text- and time-series data. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL `https://icml.cc/virtual/2026/poster/65261`.

Hyung-Chul Lee, Yoonsang Park, Soo Bin Yoon, Seong Mi Yang, Dongnyeok Park, and Chul-Woo Jung. VitalDB, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data*, 9(1):279, 2022. doi: 10.1038/s41597-022-01411-5.

Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D. Salim. SensorLLM: Aligning large language models with motion sensors for human activity recognition. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 354–379. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.19.

Zengding Liu, Chen Chen, Jiannong Cao, Minglei Pan, Jikui Liu, Nan Li, Fen Miao, and Ye Li. Large language models for cuffless blood pressure measurement from wearable biosignals. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–11. Association for Computing Machinery, 2024. doi: 10.1145/3698587.3701447.

Yunfei Luo, Yuliang Chen, Asif Salekin, and Tauhidur Rahman. Toward foundation model for multivariate wearable sensing of physiological signals. *arXiv preprint arXiv:2412.09758*, 2024. doi: 10.48550/arXiv.2412.09758. URL `https://arxiv.org/abs/2412.09758`.

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(4):1689–1696, 2021. doi: 10.3758/s13428-020-01516-y.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. IMU2CLIP: Language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13246–13253. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.883.

Guangkun Nie, Xiaocheng Fang, Gongzheng Tang, Yujie Xiao, Jun Li, Bo Liu, Hongyan Li, and Shenda Hong. Anyppg: An ecg-guided ppg foundation model trained on over 100,000 hours of recordings for holistic health profiling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 11717–11727, 2026.

OpenAI. GPT-5.6 Luna model. https://developers.openai.com/api/docs/models/gpt-5.6-luna, 2026a. OpenAI API documentation. Accessed September 25, 2026.

OpenAI. GPT-6 Luna model. https://developers.openai.com/api/docs/models/gpt-6-luna, 2026b. OpenAI API documentation. Accessed September 25, 2026.

OpenAI. GPT-6 Sol model. https://developers.openai.com/api/docs/models/gpt-6-sol, 2026c. OpenAI API documentation. Accessed September 25, 2026.

Hung Manh Pham, Jinyang Wu, Xiao Ma, Yiming Zhang, Yixin Xu, Aaqib Saeed, Bin Zhu, Zhou Pan, and Dong Ma. PulseLM: A foundation dataset and benchmark for PPG-text learning. *arXiv preprint arXiv:2603.03331*, 2026. doi: 10.48550/arXiv.2603.03331.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. PaPaGei: Open foundation models for optical physiological signals. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kYwTmlq6Vn.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Mithun Saha, Maxwell A. Xu, Wanting Mao, Sameer Neupane, James M. Rehg, and Santosh Kumar. Pulse-PPG: An open-source field-trained PPG foundation model for wearable applications across lab and field settings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 9(3):1–35, 2025. doi: 10.1145/3749494.

Satya Narayan Shukla and Benjamin M. Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

Pragya Singh, Ritvik Budhiraja, Ankush Gupta, Anshul Goswami, Mohan Kumar, and Pushpendra Singh. EEVR: A dataset of paired physiological signals and textual descriptions for joint emotion representation learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 15765–15778. Curran Associates, Inc., 2024. doi: 10.52202/079017-0503.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33536–33549. Curran Associates, Inc., 2022. doi: 10.52202/068431-2430.

Fuying Wang, Jiacheng Xu, and Lequan Yu. From token to rhythm: A multi-scale approach for ECG-language pretraining. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 65059–65074. PMLR, 2025.

Xiaoda Wang, Ching Chang, Defu Cao, Kaiqiao Han, Fang Sun, Yue Huang, Minxiao Wang, Chang Xu, Xiao Luo, Runze Yan, Xiangliang Zhang, Xiao Hu, Yan Liu, Yizhou Sun, Wei Wang, and Carl Yang. Position: Beyond prediction: Toward verifiable physiological waveform reasoning with foundation models and agentic LLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/67119.

Yiping Xie, Bo Zhao, Mingtong Dai, Jian-Ping Zhou, Yue Sun, Tao Tan, Weicheng Xie, Linlin Shen, and Zitong Yu. PhysLLM: Harnessing large language models for cross-modal remote physiological sensing. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=aR43t8OEeW.

Zongzhe Xu, Zitao Shuai, Eideen Mozaffari, Ravi S. Aysola, Rajesh Kumar, and Yuzhe Yang. SleepLM: Natural-language intelligence for human sleep. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/65807.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=Ee277P3AYC.

Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. HecVL: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15006 of *Lecture Notes in Computer Science*, pp. 306–316. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72089-5_29.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A Ali Heydari, Girish Narayanswamy, Max Xu, Ahmed Metwally, Jinhua Xu, Jake Garrison, Xuhai Xu, et al. Sensorlm: Learning the language of wearable sensors. *Advances in Neural Information Processing Systems*, 38:46813–46849, 2026a.

Zexing Zhang, Huimin Lu, and Qingxin Zhao. PPGPT: Transferring next-token modeling from language to PPG signals. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(34): 28573–28581, 2026b. doi: 10.1609/aaai.v40i34.40088.

A LIMITATIONS

We pretrain PPG-LM on single-center adult emergency-department data. Although MIMIC-III and VitalDB provide external validation, broader multi-center data would strengthen generalization across clinical settings. Clinical supervision depends on available EHR information, whose completeness and temporal accuracy affect caption quality and may limit model understanding and generalization. Richer longitudinal records and clinician-verified descriptions could improve this supervision. Medication-event recognition captures associations rather than causal drug effects, since treatment indication and concurrent interventions may confound observed signals. The current model is intended for research and is not ready for direct clinical decision-making; prospective validation, clinician assessment, subgroup analyses, and uncertainty calibration are needed before clinical deployment.

B CAPTION CONSTRUCTION DETAILS

We summarize how waveform measurements and clinical records produce segment, visit, event, and temporal captions. The tables describe their content; Appendix E provides language templates and rendered examples.

B.1 SEGMENT CAPTIONS AND SIGNAL QUALITY

Measurements. Each segment contains 30 seconds of PPG sampled at 125 Hz. NeuroKit2 provides pulse timing, amplitude, and quality measurements; pyPPG provides augmentation and signal-quality indices (Makowski et al., 2021; Goda et al., 2024). Synchronized ECG supplies rhythm annotations and supports quality checks; the model input remains PPG.

Table 9: **Segment-caption content.** Perfusion and morphology are waveform surrogates rather than independent clinical diagnoses.

Attribute	Evidence and description
Pulse rate	Median pulse-interval rate, expressed as a value and a category.
Rate trend	Instantaneous rate slope: falling, stable, or rising.
Rhythm	ECG R–R interval variability, expressed as a regularity category.
Perfusion	Pulsatile amplitude relative to baseline: weak, moderate, or strong.
Signal quality	NeuroKit2 quality score: clean, fair, or noisy but usable.
Augmentation	The pyPPG augmentation index: low, moderate, or high.
SVRI	Post-peak/pre-peak mean-amplitude ratio in normalized beats: low, mid-range, or high.
Amplitude variability	Beat-amplitude variation: steady to highly variable.

Quality filtering. Usability combines PPG quality with available signal-quality, flatline, and clipping checks; missing optional checks do not reject a window. Good windows meet stricter criteria, including PPG–ECG rate agreement when reliable ECG exists; other usable windows are fair. Selection tiers differ from caption quality labels. Unavailable or unreliable ECG rhythm annotations are omitted.

Rendering. Missing features are omitted. Poor windows receive quality-only descriptions and are excluded from training but retained for validation and testing. Each feature has ten templates; template selection and sentence shuffling use reproducible example-specific seeds.

Signal-plus-vitals view. For the signal-plus-vitals configuration reported in Section 5, available charted HR, SpO_2, RR, and BP measurements augment the eight signal attributes described in Section 3. Each value is matched to the nearest charted observation within 300 seconds of the window timestamp; unavailable values are omitted. These measurements enter the text target, while the waveform input remains PPG.

B.2 VISIT CAPTIONS AND CLINICAL CONTEXT

Clinical records. Visit captions summarize encounter context (Table 10). History entries are deduplicated and ranked by clinical relevance, then recency.

Table 10: **Visit-caption content.**

Component	Evidence and description
Demographics	Recorded age and sex.
Presentation	Arrival mode, Emergency Severity Index, and normalized chief complaints.
Triage measurements	Available heart rate, respiratory rate, oxygen saturation, blood pressure, and temperature.
Clinical history	Selected prior ICD-10 diagnoses, expressed as clinical descriptions.
Home medications	Recorded active home-medication classes.

Field selection. Known post-arrival diagnoses are excluded; history with missing diagnosis or arrival dates is retained. Home medications use patient-level active records without encounter-specific temporal verification. These fields therefore have incomplete evidence of availability time. Disposition, discharge diagnoses, and length of stay are excluded.

B.3 EVENT CAPTIONS AND WITHIN-VISIT CONTROLS

Administration records. Eligible records require a parseable first-administration timestamp and a normalized drug name. Route and class filters exclude topical, ophthalmic, and other ineligible administrations. Name rules and class-keyword fallback map the remaining records to opioids, intravenous fluids, rate-control agents, vasodilators, bronchodilators, sedatives, non-opioid analgesics/antipyretics, vasopressors, diuretics, or antiemetics. Unmapped records do not become targets; near-duplicate administrations are removed within each visit and class.

Table 11: **Event-caption construction.**

Component	Evidence and description
Event anchor	A recorded administration timestamp for an eligible medication class.
Pre-event context	The latest available PPG windows preceding the administration.
Post-event context	PPG windows sampled evenly across the period following administration.
Caption target	The medication class, with individual drug names, doses, and exact latencies omitted.
Control view	The same sampling rule around a within-visit anchor free of nearby eligible administrations.

Sampling and controls. Event views use up to three windows in the preceding 30 minutes and up to five in the following 60 minutes, with at least one pre-administration and two post-administration windows. Control anchors are sampled uniformly over the same visit's monitored span; their exclusion interval also covers eligible but unmapped drugs. Controls are capped at the retained event count, subject to availability. This construction avoids a fixed window-start cue but does not establish causal effects or remove all temporal confounding. Antiemetics serve as an additional comparison class.

B.4 TEMPORAL DESCRIPTIONS AND CORRUPTED STATEMENTS

Trajectory evidence. Ordered segment-caption rates and rhythm labels produce the temporal descriptions in Table 12. They summarize observed tachycardia, irregularity, and changes in pulse rate across a visit.

Table 12: **Temporal-caption content.**

Operator	Evidence and description
Occurrence	Whether tachycardia or irregularity appears among known observations.
Proportion	The affected share of known windows, described as brief, part, or most.
Contiguous duration	Consecutive affected windows, grouped into short, intermediate, or long runs.
Rate ordering	Consecutive groups of observed rates, describing rising, falling, stable, rise-then-fall, or fall-then-rise patterns.

Missing observations. Proportion and duration descriptions require the state to occur. Unknown states and monitoring gaps interrupt runs; unsupported operators and ambiguous rate profiles are omitted.

Corrupted descriptions. Each supported operator value has three templates. A negative substitutes an incompatible value in one sentence and re-renders that sentence, retaining the remaining facts. Padded sentences and unavailable negatives are masked.

C METHOD AND TRAINING DETAILS

This appendix expands the architecture and alignment objectives in Section 4, using the same notation throughout. The waveform feature is g_i, the level-conditioned text embedding is $u^\ell(c)$, and norm denotes ℓ_2 normalization. We first specify the two encoders and their CoCa training interface, then describe event and visit aggregation, preservation of segment alignment, and the two-stage optimization schedule.

C.1 WAVEFORM AND TEXT ENCODERS

Waveform encoder. The encoder f_θ operates on a single-channel, 30-second PPG window $x_i \in \mathbb{R}^{3750}$ sampled at 125 Hz. Each window is standardized using its own mean and standard deviation, with 10^{-6} added to the denominator. A one-dimensional convolution with kernel and stride 25 produces 150 non-overlapping patches, each spanning 0.2 seconds. Learned absolute positional embeddings are added to the width-384 patch embeddings before six Transformer blocks (Vaswani et al., 2017), each with six attention heads and an MLP hidden width of 4×384. Denoting the final normalized patch states by $H_i \in \mathbb{R}^{150\times 384}$, the two waveform representations are

$$g_i = \mathrm{Pool}_1(H_i) \in \mathbb{R}^{384}, \qquad z_i^{\mathrm{seg}} = \mathrm{norm}(W_{\mathrm{seg}} g_i) \in \mathbb{R}^{512}. \tag{6}$$

Here Pool_1 is a single learned-query attention pooler with a residual MLP, and $W_{\mathrm{seg}} \in \mathbb{R}^{512\times 384}$ is bias-free. The waveform encoder, including this projection, has approximately 12.7M parameters. As in the clinical prediction experiments, the 768-dimensional window probe feature concatenates g_i with the mean of the 150 patch states, before the alignment projection.

Text encoder. The shared text tower ψ_ϕ is an eight-layer causal Transformer with width 512, eight attention heads, and an MLP hidden width of 4×512. It uses CLIP's byte-pair encoding vocabulary of 49,408 tokens (Radford et al., 2021), learned token and absolute positional embeddings, and a learned prefix e_ℓ for $\ell \in \{\mathrm{seg}, \mathrm{evt}, \mathrm{vis}\}$. The prefix is prepended to the text sequence and participates in causal self-attention. The level-conditioned read-out in Section 4 is

$$u^\ell(c) = \mathrm{norm}(P_\ell \psi_\phi([e_\ell; c])_{\mathrm{EOS}}), \qquad P_\ell \in \mathbb{R}^{512\times 512}. \tag{7}$$

The end-of-sequence state summarizes the available text, and the level-specific projection selects its alignment space. The same tower encodes full captions and individual temporal statements; the latter use the visit prefix and projection. Both waveform and text towers are trained from random initialization. CLIP supplies the tokenizer and contrastive formulation, rather than pretrained encoder weights.

C.2 COCA ALIGNMENT AND CAPTIONING OBJECTIVES

Contrastive branch. CoCa combines paired representation learning with conditional caption generation (Yu et al., 2022). In our waveform adaptation, the contrastive branch connects the single-query waveform read-out in Equation 6 to the text-only end-of-sequence read-out in Equation 7. For a batch of B matched window–caption pairs, the normalized embeddings give

$$\begin{aligned} Q_{ij} &= s_{\mathrm{seg}} (z_i^{\mathrm{seg}})^\top u^{\mathrm{seg}}(c_j), \qquad \mathcal{C}_i^{\mathrm{seg}} = \{i\} \cup \{j : \kappa_j \neq \kappa_i\}, \\ p_{ij}^{\mathrm{seg}\to\mathrm{text}} &= \frac{\exp Q_{ij}}{\sum_{k\in\mathcal{C}_i^{\mathrm{seg}}} \exp Q_{ik}}, \qquad p_{ij}^{\mathrm{text}\to\mathrm{seg}} = \frac{\exp Q_{ji}}{\sum_{k\in\mathcal{C}_i^{\mathrm{seg}}} \exp Q_{ki}}, \quad j \in \mathcal{C}_i^{\mathrm{seg}}. \end{aligned} \tag{8}$$

The scale $s_{\mathrm{seg}} > 0$ is learned. The content key κ_i represents the binned measurements underlying a caption, independently of template wording and sentence order. Off-diagonal pairs with equal keys are masked in both directions, while each matched pair is retained. Symmetric InfoNCE (Radford et al., 2021) minimizes the negative log-probability of the diagonal match in both directions:

$$\mathcal{L}_{\mathrm{NCE}}(Q; \mathcal{C}^{\mathrm{seg}}) = -\frac{1}{2B} \sum_{i=1}^{B} \left(\log p_{ii}^{\mathrm{seg}\to\mathrm{text}} + \log p_{ii}^{\mathrm{text}\to\mathrm{seg}} \right). \tag{9}$$

This is Equation 1 with the segment candidate set from Section 4.1. Excluded pairs do not contribute to either denominator; they are not treated as additional positives.

Captioning branch. A separate attention pooler with 32 learned queries reads the same patch states H_i and produces waveform memory $M_i = \mathrm{Pool}_{32}(H_i) \in \mathbb{R}^{32\times 384}$. A learned projection maps the memory to width 512. The causal token states from the eight-layer text tower, with the level-prefix position removed, then pass through four additional multimodal decoder layers of width 512, eight heads, and MLP ratio four. Each layer combines causal text self-attention with cross-attention to the projected waveform memory. This implements CoCa's separation between text-only contrastive features and waveform-conditioned caption features (Yu et al., 2022).

Let $d_{i,r} \in \mathbb{R}^{512}$ be the final layer-normalized decoder state that predicts token $c_{i,r}$ from the preceding caption tokens, and let $E_{\mathrm{tok}} \in \mathbb{R}^{49408\times 512}$ be the shared token embedding matrix. The tied output head and training objective are

$$\begin{aligned} p_{\phi,\eta}(c_{i,r} \mid c_{i,<r}, M_i) &= [\mathrm{softmax}(E_{\mathrm{tok}} d_{i,r})]_{c_{i,r}}, \\ \mathcal{L}_{\mathrm{cap}} &= -\frac{1}{N_{\mathrm{tok}}} \sum_i \sum_{r\in\mathcal{T}_i} \log p_{\phi,\eta}(c_{i,r} \mid c_{i,<r}, M_i), \\ \mathcal{L}_{\mathrm{seg}} &= \mathcal{L}_{\mathrm{NCE}}(Q; \mathcal{C}^{\mathrm{seg}}) + \mathcal{L}_{\mathrm{cap}}. \end{aligned} \tag{10}$$

Here $\mathcal{T}_i$ contains non-padding target positions, including EOS, $N_{\mathrm{tok}} = \sum_i |\mathcal{T}_i|$, and η denotes the additional parameters of the waveform-memory and multimodal decoding branch, as in Section 4.1. Training uses teacher forcing and unit weights for both losses. The two branches share the waveform patch encoder and causal text tower, so caption supervision also updates the representations used for alignment. Caption loss applies to segment descriptions in every batch type; event and visit descriptions are alignment targets. The CLIP objective compared in the experiments retains the contrastive branch without caption loss.

C.3 Event and visit aggregation

Event read-out. For a medication or control anchor a_j, E_j contains up to three latest windows in $[a_j - 30, a_j)$ and up to five evenly spaced windows in $[a_j, a_j + 60]$, following Appendix B.3. Both t_i and a_j use the visit clock, so $t_i - a_j$ is a signed offset in minutes. Each view has at least one pre-anchor and two post-anchor windows. The time embedding φ maps 64 sinusoidal features, with periods from two minutes to two weeks, through an MLP to width 384. The attention operator in Equation 2 includes layer normalization of the waveform content and is implemented by

$$\begin{aligned} U_j^{(0)} &= [q_{\mathrm{evt}}; (\mathrm{LN}(g_i) + \varphi(t_i - a_j))_{i\in E_j}], \\ U_j^{(b)} &= \mathcal{B}_b^{\mathrm{evt}}(U_j^{(b-1)}), \quad b = 1, 2, \\ z_j^{\mathrm{evt}} &= \mathrm{norm}\Big(W_{\mathrm{evt}}\, \mathrm{LN}([U_j^{(2)}]_0)\Big). \end{aligned} \tag{11}$$

Each $\mathcal{B}_b^{\mathrm{evt}}$ is a self-attention block of width 384, six heads, and MLP ratio four; key-padding masks exclude missing windows. Reading the learned-query position yields one 512-dimensional event embedding. Signed offsets preserve the distinction between observations before and after the anchor, including for control views.

Visit read-out and reference times. For a visit V, we use the input construction in Section 4.2:

$$b_i = g_i + \gamma_{\mathrm{mon}}(t_i) + \gamma_{\mathrm{arr}}(\delta_i), \qquad B_V = [(b_i)_{i\in\mathcal{I}_V}; G_V; e_V]. \tag{12}$$

Here $\mathcal{I}_V$ indexes at most 256 selected windows, G_V contains gap tokens, and e_V is the evidence token. We prepend eight learned latent queries and update the complete sequence through two self-attention blocks of width 384, six heads, and MLP ratio four:

$$\begin{aligned} U_V^{(0)} &= [q_0; \ldots; q_7; B_V], \\ U_V^{(b)} &= \mathcal{B}_b^{\mathrm{vis}}(U_V^{(b-1)}; \tau_V), \quad b = 1, 2, \\ \ell_{Vk} &= \mathrm{LN}([U_V^{(2)}]_k), \qquad h_{Vk} = \mathrm{norm}(W_{\mathrm{lat}}\ell_{Vk}), \quad k = 0, \ldots, 7. \end{aligned} \tag{13}$$

The blocks use the token times τ_V and a validity mask; both the latent and observation positions are updated. Inspired by Perceiver's learned latent representations (Jaegle et al., 2021), only the eight latent positions are read out. Attention operates on the full concatenated sequence, with at most

$8 + 256 + 16 + 1 = 281$ tokens. The global latent has reference time $r_0 = |\mathcal{I}_V|^{-1} \sum_{i \in \mathcal{I}_V} t_i$ and supplies $z_V^{\text{vis}} = h_{V0}$ for clinical-caption alignment. The remaining reference times are $(r_1, \ldots, r_7) = (0, 15, 30, 60, 120, 240, 480)$ minutes after the first window. Following the reference-time idea in mTAN (Shukla & Marlin, 2021), these anchors give the queries temporal identities while allowing attention to all valid observations.

Continuous-time rotary encoding. We adapt rotary position embeddings from RoFormer (Su et al., 2024) to elapsed minutes. In each 64-dimensional attention head, 16 channel pairs are rotated and the other 32 channels are unchanged. With $T_p = 2(20160/2)^{(p-1)/15}$ minutes and $\theta_p = 2\pi/T_p$, define

$$R_\Theta(t) = \operatorname{diag}_{p=1}^{16} \begin{pmatrix} \cos(\theta_p t) & -\sin(\theta_p t) \\ \sin(\theta_p t) & \cos(\theta_p t) \end{pmatrix}. \tag{14}$$

For query and key times τ and τ', respectively,

$$\begin{aligned} \widetilde{q}(\tau) &= [R_\Theta(\tau) q^{\text{rot}}; q^{\text{free}}], & \widetilde{k}(\tau') &= [R_\Theta(\tau') k^{\text{rot}}; k^{\text{free}}], \\ \widetilde{q}(\tau)^\top \widetilde{k}(\tau') &= (q^{\text{rot}})^\top R_\Theta(\tau' - \tau) k^{\text{rot}} + (q^{\text{free}})^\top k^{\text{free}}. \end{aligned} \tag{15}$$

Setting $\tau = r_k$ and $\tau' = t_i$ recovers Equation 3. Each head divides these scores by $\sqrt{64}$, applies the validity mask, and normalizes with softmax. Using actual minutes distinguishes irregular intervals that share the same window-index separation. The relative-rotation identity holds for fixed query/key content; additive clock features also affect that content.

Time content and monitoring coverage. As in Equation 12, learned MLPs map 64-dimensional sinusoidal features of monitoring time t_i and time since ED arrival δ_i into additive width-384 features. A learned vector represents unavailable arrival time. These features expose time through the values as well as the attention scores. Gaps longer than ten minutes are detected over the full visit before context subsampling; the longest 16 are retained. Each gap token uses its midpoint as its time and encodes the logarithmically transformed duration, $\log(1 + \text{duration})$. The evidence token maps six coverage descriptors through a $6 \to 384 \to 384$ MLP: window count, coverage fraction, monitoring span, fraction of good-quality windows, arrival-to-monitoring delay, and an arrival-time availability indicator. Window count, monitoring span, and available arrival delay use $\log(1+\cdot)$ transforms, with times measured in minutes. Coverage is $\min(1, 0.5n/\max(\text{span}, 0.5))$ for n observed 30-second windows, with a half-minute lower bound on the denominator for a single-window visit. These inputs describe the amount and timing of observed evidence without imputing physiology within monitoring gaps.

Alignment across scopes. Event alignment uses Equation 1 with $Q_{ij} = s_{\text{evt}}(z_i^{\text{evt}})^\top u^{\text{evt}}(c_j)$ and $\mathcal{C}_i^{\text{evt}} = \{i\} \cup \{j : \text{class}(j) \neq \text{class}(i)\}$; all control views share one category. Global visit alignment uses $Q_{ij} = s_{\text{vis}}(z_i^{\text{vis}})^\top u^{\text{vis}}(c_j)$ with the full in-batch candidate set. Temporal matching uses all eight h_{Vk} and the mean-over-statements, max-over-latents score R_{ij} in Equation 4.2, adapting ColBERT's late interaction (Khattab & Zaharia, 2020) to whole statements. Each visit supplies at most eight statements, with no fixed statement-to-latent assignment. Equation 4 combines in-batch alignment of $s_{\text{vis}} R$ with a corruption ranking term of weight 0.5. For the temporal contrastive term, we retain the matched sub-batch of visits with $m_V > 0$ and use all candidates within that sub-batch; the term is zero if no visit has available statements. Padded statements do not enter the mean defining R_{ij}. The ranking term averages only over visits with an available corruption; if none are available, this term is zero. A visit without temporal statements still participates in clinical-caption alignment. Appendix B.4 specifies the corruptions. No cross-batch negative queue is used at any scope.

C.4 Preserving segment alignment

Stage 2 introduces event and visit targets whose temporal scope differs from that of a segment caption. Updating the shared encoder solely for these broader targets can change the local waveform–language geometry learned in Stage 1. We therefore retain the segment objective on every batch and use a frozen Stage-1 waveform encoder as a reference, alongside the multilevel training strategy motivated by HierVL (Ashutosh et al., 2023). The ongoing segment objective continues to learn from paired captions; the reference penalty limits changes to the normalized waveform embeddings.

Let $\mathcal{Q}$ be the windows re-encoded in the current batch. A frozen copy of the Stage-1 waveform encoder and segment projection supplies $z_i^{(1)} = \text{norm}(W_{\text{seg}}^{(1)} g_i^{(1)})$. With no gradients through this

reference branch, we use

$$\begin{aligned}\mathcal{L}_{\mathrm{dst}}(\mathcal{Q}) &= \frac{1}{|\mathcal{Q}|}\sum_{i\in\mathcal{Q}}\left(1-(z_i^{\mathrm{seg}})^\top z_i^{(1)}\right)\\ &= \frac{1}{2|\mathcal{Q}|}\sum_{i\in\mathcal{Q}}\left\|z_i^{\mathrm{seg}}-z_i^{(1)}\right\|_2^2.\end{aligned} \tag{16}$$

The equality follows from unit normalization. The penalty constrains embedding directions while allowing the pre-projection waveform features to adapt. The teacher remains in evaluation mode with dropout disabled; the current waveform encoder and projection receive gradients after the initial freezing period in Appendix C.5.

The alignment terms in Equation 5 can then be written as

$$\begin{aligned}\mathcal{L}^{(\mathrm{seg})} &= \mathcal{L}_{\mathrm{seg}}(\mathcal{Q}) + \lambda_{\mathrm{dst}}\mathcal{L}_{\mathrm{dst}}(\mathcal{Q}),\\ \mathcal{L}^{(\mathrm{evt})} &= \mathcal{L}_{\mathrm{seg}}(\mathcal{Q}) + \lambda_{\mathrm{dst}}\mathcal{L}_{\mathrm{dst}}(\mathcal{Q}) + \mathcal{L}_{\mathrm{evt}},\\ \mathcal{L}^{(\mathrm{vis})} &= \mathcal{L}_{\mathrm{seg}}(\mathcal{Q}) + \lambda_{\mathrm{dst}}\mathcal{L}_{\mathrm{dst}}(\mathcal{Q}) + \mathcal{L}_{\mathrm{vis}} + \mathcal{L}_{\mathrm{tmp}},\end{aligned} \tag{17}$$

where $\lambda_{\mathrm{dst}} = 0.1$. For segment and event batches, $\mathcal{Q}$ contains their valid input windows, each paired with its own segment caption. For visit batches, it contains up to eight live windows per visit; detached context features participate in aggregation but incur no segment or distillation loss. The same content-key mask is applied whenever the segment objective is evaluated, including on live windows from event and visit batches.

C.5 Optimization, feature caching, and inference

Two-stage schedule. Stage 1 uses the segment objective in Section 4.1. Both towers are trained from scratch for six epochs over 8,842,271 training windows, using batches of 256 and dropping the final incomplete batch. This gives 207,240 updates. Stage 2 initializes from Stage 1 and trains the shared encoders and scope-specific read-outs under the freezing schedule below for 30,000 global updates, cycling through segment, visit, and event batches in a 4:2:1 ratio. Batch sizes are 256 windows, 64 visits, and 64 event views, respectively. The extended Stage-2 run used for the training-dynamics analysis in Figure 4(b) is separate from this standard schedule.

Optimization. Both stages use AdamW with $\beta = (0.9, 0.98)$, $\epsilon = 10^{-6}$, peak learning rate 5×10^{-4}, and weight decay 0.2. The learning rate follows a 2,000-update linear warmup and cosine decay to zero over each stage's own schedule. Weight decay excludes biases, normalization parameters, token and positional embeddings, learned queries and prefixes, missing-time embeddings, and logit scales. During the first 2,000 Stage-2 updates, the waveform and text towers and the segment logit scale have zero learning rate; subsequently they use one tenth of the scheduled rate. Aggregators and other read-out parameters use the full scheduled rate from the start. Global gradient norms are clipped at 1.0. Dropout is 0.1 in waveform Transformer blocks and the event and visit aggregators, and zero in the text tower, caption decoder, and attention poolers. Each level has a learned scale $s_\ell = \min(\exp(\alpha_\ell), 100)$; newly introduced scales are initialized to $1/0.07$, while Stage 2 retains the learned segment scale. Token budgets are 256 for segment and visit captions, 64 for event captions, and 48 for temporal statements.

Loss weights. All contrastive alignment terms and the segment captioning loss use unit weights. Stage 2 uses a distillation weight of 0.1 and a temporal corruption-ranking weight of 0.5.

Live windows and cached context. The Stage-1 waveform encoder initializes an fp16 cache of the 384-dimensional window features g_i. Each visit batch selects a context of at most 256 windows per visit. Up to eight distinct windows per visit are sampled for live re-encoding, using all windows when fewer than eight are available; their features replace the corresponding cached entries within this context, so the live windows count toward the context limit. For the selected set $\mathcal{I}_V$ and live subset $\mathcal{Q}_V = \mathcal{Q} \cap \mathcal{I}_V$, aggregation uses

$$\widetilde{g}_i = \begin{cases} f_\theta(x_i), & i \in \mathcal{Q}_V,\\ \mathrm{stopgrad}(g_i^{\mathrm{cache}}), & i \in \mathcal{I}_V \setminus \mathcal{Q}_V. \end{cases} \tag{18}$$

This lets visit-level losses update the aggregators over a longer context while limiting waveform-encoder activations to the live subset. The context cap at global Stage-2 update s is

$$C_{\max}(s) = \left\lfloor 8 + 248 \min\left(1, \frac{s}{10{,}000}\right) \right\rfloor . \tag{19}$$

Half of the visit batches use this cap; the other half draw a context budget log-uniformly between 8 and the current cap, subject to the available windows. Longer visits are subsampled uniformly over time while retaining the live windows. Detached live features are written back to the cache after visit batches, and segment batches also refresh their encoded windows. Event views are encoded directly from raw windows and do not update the cache. Each training run uses its own cache.

Inference. All model parameters are frozen. We encode the windows selected by each benchmark using the evaluation checkpoint; the training feature cache is not required. Segment and event tasks use z^{seg} and z^{evt}, respectively, while clinical-caption visit retrieval uses the global latent $z_V^{\text{vis}} = h_{V0}$. Retrieval and class-prompt scoring compare normalized waveform and corresponding level-conditioned text embeddings by cosine similarity; task-specific prompt sets and read-outs are specified in Appendix D. Temporal descriptions use all eight latents and the score in Equation 4.2. Segment captioning starts with a beginning-of-sequence token and uses greedy autoregressive decoding conditioned on M_i, stopping at EOS or the segment token limit.

Compute. The main CoCa model is trained with seed 0 and bfloat16 autocast on one NVIDIA H200 GPU. Recorded training times are approximately 26.3 hours for Stage 1 and 6.0 hours for Stage 2, excluding cache construction and evaluation. Caption cross-entropy is computed in chunks of 8,192 target tokens with gradient checkpointing to reduce memory use, without changing the token-averaged objective in Equation 10.

D Additional Evaluation Details

D.1 Experimental setup

Datasets and patient splits. The three datasets cover emergency, intensive-care, and intraoperative monitoring. Only MC-MED is used to pretrain PPG-LM; MIMIC-III and VitalDB evaluate transfer to external clinical settings.

- **MC-MED** (Kansal et al., 2025) pairs PPG from Stanford's adult emergency department (ED) with structured electronic health records. The processed cohort contains 36,430 patients and 53,473 visits. The patient-random 80/10/10 split yields the retained counts in Table 13. PPG is divided into non-overlapping 30-second windows at 125 Hz; the training set provides approximately 73,700 hours. Poor-quality windows are excluded from pretraining, while the validation and test corpus splits retain them.
- **MIMIC-III** (Johnson et al., 2016) links intensive-care bedside waveforms from Beth Israel Deaconess Medical Center to clinical records. The external benchmark uses PPG paired with synchronized ECG where available, charted vital signs, demographics, laboratory measurements, and hospital outcomes. Language evaluation uses fixed patient-level samples. Clinical probes are fitted on the target dataset's training patients and evaluated on patient-disjoint test patients.
- **VitalDB** (Lee et al., 2022) pairs intraoperative biosignals from Seoul National University Hospital with perioperative clinical information. It provides preoperative conditions and laboratory measurements, intraoperative monitoring, medication records, and postoperative ICU outcomes. The language benchmark has official 40 Hz and original 500 Hz waveform views of the same evaluation items. The main PPG-LM language results, retrieval analysis, and prompt-sensitivity analysis use the 500 Hz source; supplementary caption/read-out and few-shot analyses use the 40 Hz view. Baselines retain their recorded input views. Clinical probes use target-dataset training patients and patient-disjoint testing.

Waveform preparation and references. PPG-LM inputs are resampled to 125 Hz and standardized within each window as $(x - \mu)/(\sigma + 10^{-6})$. The captioning pipeline supplies reference descriptions on all three datasets; Appendix B describes its signal measurements, quality gates, and

Table 13: **Processed MC-MED splits.** Patients are disjoint across splits; window counts refer to the retained corpus.

Split	Patients	Visits	30-second windows
Training	29,177	42,856	8,842,271
Validation	3,641	5,421	1,183,278
Test	3,612	5,196	1,099,395
Total	36,430	53,473	11,124,944

clinical-record matching. ECG supports rhythm annotation and the independent HR reference, but is not a model input.

Baseline configurations. For language tasks, GPT-5.6-luna, GPT-6-luna, GPT-6-sol, and Claude Opus 5.5 receive each 30-second PPG window mean-pooled to 25 Hz and standardized, giving 750 numerical values, and return JSON answers. Prompts include the annotation class definitions without reference measurements. Signal-to-text retrieval shuffles candidate-caption order independently for each query and requests five preferred indices. PulseLM uses its Qwen2.5-7B-Instruct variant. Our Qwen3-4B (SFT) baseline uses a frozen PaPaGei-S encoder and full language-model fine-tuning on 100,000 window and 40,000 visit examples from our captioning corpus; evaluation uses prompt scoring or generated answers as appropriate to the task. Clinical prediction compares frozen PaPaGei-S, Pulse-PPG, AnyPPG, SIGMA-PPG, MOMENT-L, and Chronos-2 features with the same logistic- or ridge-regression probe protocol (Appendix D.7). For visit-level clinical targets, all baselines and both PPG-LM stages receive the same four windows; baselines and Stage 1 use mean pooling, while Stage 2 uses learned visit aggregation. MOMENT-L is the largest evaluated MOMENT variant; PaPaGei-S is the morphology-aware variant. Unless otherwise stated, PPG-LM uses the caption-style template prompt bank.

Benchmark sampling and scoring. Language evaluations use fixed items and cached model responses. Window recognition and HR evaluation use 300 windows per dataset, with one window per patient, stratified by rhythm and HR class and excluding the poor-quality tier. Captioning uses the first 200 of these windows. Clinical-fact recognition in Table 6 uses 400 MC-MED visits and 300 VitalDB cases, with eight windows per case and enriched sampling of rare labels. Its thresholds are calibrated on held-out MC-MED validation visits or separate VitalDB patients, without using test labels. The broader visit-inference benchmark instead samples 200 encounters per dataset, requires at least eight windows, and stratifies by sex, age group, and outcome. Visit benchmarks retain one encounter per patient. The medication-class results in Table 5 use 393 administration views and 393 controls on MC-MED and 139 of each on VitalDB; Appendix D.6 additionally reports the smaller event cohort. Off-label or missing classification answers count as errors. Numerical MAE uses items with both a numerical answer and an available reference; the HR comparisons have 300 valid ECG-referenced items per system and dataset. Dashes denote unavailable or unreported evaluations.

Evaluation protocol. Unless otherwise stated, PPG-LM denotes the Stage-2 model. All model backbones remain frozen during evaluation. Within each benchmark, comparisons use PPG inputs at the same temporal level and with the same number of windows across models. Language benchmarks use the same evaluation examples and class definitions across models; PPG-LM performs prompt-based recognition, cross-modal retrieval using cosine similarity, and caption generation with greedy decoding. External zero-shot segment recognition, retrieval, and captioning use no target-dataset fitting. Clinical-fact recognition uses read-out thresholds calibrated on held-out MC-MED visits or separate VitalDB patients. For supervised clinical prediction, we fit logistic regression probes for classification and ridge regression probes for continuous targets on frozen features from training patients, and evaluate on patient-disjoint test sets.

D.2 Clinical task definitions

We define the physiological attributes, medication and visit facts, and clinical prediction targets used in the main experiments. Tables 14 and 15 explain the language-based tasks; Tables 16 and 17 cover all 32 clinical prediction targets (17 MC-MED, six MIMIC-III, and nine VitalDB). In these

tables, M denotes MC-MED, I denotes MIMIC-III, and V denotes VitalDB. Definitions distinguish the recorded target from its clinical interpretation.

Physiological understanding. Window-level tasks describe rate, regularity, signal properties, and charted physiology (Table 14). Their annotation rules are given in Appendix B.1. Rhythm regularity is a window attribute, whereas atrial fibrillation in the clinical probe is a recorded prior diagnosis. Cross-modal retrieval tests whether a signal retrieves its paired description and vice versa; R@k measures whether that pair appears in the top k candidates (Appendix D.3). It is a correspondence task rather than a separate disease label.

Table 14: **Physiological task definitions.** Tasks in Tables 2 and 3, including the HR-class target used in few-shot evaluation. CV denotes coefficient of variation.

Task	Reference and clinical interpretation
Rhythm	Regular, slightly irregular, or markedly irregular according to synchronized ECG R–R interval CV (< 0.05, $[0.05, 0.12)$, or ≥ 0.12). Measures beat-timing regularity, without assigning a specific arrhythmia.
Quality (Qual.)	Clean, fair, or noisy according to the annotation quality score and usability gate. Describes waveform reliability and artifact burden.
Perfusion (Perf.)	Weak, moderate, or strong pulsatile amplitude relative to the signal baseline, using the waveform-derived AC/DC index. A peripheral perfusion surrogate, not a direct blood-flow measurement.
Morphology (Morph.)	Low, moderate, or high augmentation from the pyPPG augmentation index. Describes pulse-wave shape; it does not establish a vascular disease diagnosis.
HR class	Five pulse-rate bins: < 50, 50–59, 60–100, 101–120, and > 120 bpm, spanning markedly slow to fast rates. These are benchmark annotation categories.
HR estimation	Heart rate in beats/min, evaluated against synchronized ECG. This independent reference differs from the PPG-derived pulse rate used to construct training captions.
RR class / estimation	Low, normal, or elevated respiratory rate, or its numerical value in breaths/min. Both use charted respiratory rate as the reference and describe breathing frequency.
SBP / MAP	Systolic blood pressure (peak arterial pressure during a beat) and mean arterial pressure (pressure averaged over the cardiac cycle), in mmHg, evaluated against the corresponding charted measurements.
HR trend	Falling, stable, or rising rate within a window, from the slope of instantaneous pulse rate. Included in caption factuality.
Amplitude variability	Steady, moderately variable, or highly variable beat amplitudes, from beat-amplitude CV. Included in caption factuality.
Caption factuality (Fact)	Pooled accuracy of stated facts across six shared fields: HR, HR class, rhythm, HR trend, amplitude variability, and quality. Numerical HR is correct within ± 5 bpm of the caption reference; categorical fields require agreement.

Medication events and visit facts. Table 15 defines every medication class and binary visit fact displayed in the main results. Event AUROC compares each medication class with no-administration controls. The class describes the recorded administration, without asserting a causal physiological response. Home-medication facts instead use patient-level active medication records, whose timing is not verified for each encounter (Appendix B.2); their classes may overlap. VitalDB facts describe preoperative status. In particular, the binary eGFR target identifies reduced estimated filtration, without establishing the duration required for a chronic kidney disease diagnosis.

Table 15: **Medication and visit-fact definitions.** Dataset abbreviations refer to the tasks displayed in Tables 5 and 6.

Task	Data	Recorded target and clinical meaning
Administration events: medication class versus control		
Rate control	M	Administration of a rate-controlling medication, used to slow heart rate.
Vasopressor	M, V	Administration of an agent used to increase arterial blood pressure.
Vasodilator	M, V	Administration of a medication that dilates blood vessels.
IV fluid	M	Intravenous fluid administration, representing fluid delivery to the circulation.
Diuretic	M	Administration of medication that promotes renal salt and water excretion.
Opioid	V	Administration of an opioid analgesic, used for pain control.
Sedative	V	Administration of medication used to induce sedation or reduce arousal.

Table 15 (continued)

Task	Data	Recorded target and clinical meaning
Recorded home-medication use: binary visit facts		
CCB	M	Recorded use of calcium-channel blockers, which act on vascular and/or cardiac muscle.
AntiHTN	M	Recorded use of antihypertensive medication, used to lower blood pressure.
AntiPL	M	Recorded use of antiplatelet medication, which inhibits platelet activation or aggregation.
AntiDM	M	Recorded use of glucose-lowering medication for diabetes management.
Preoperative status: binary visit facts		
HTN	V	Preoperative hypertension, a recorded condition of persistently elevated blood pressure.
DM	V	Preoperative diabetes mellitus, a disorder of glucose regulation.
Anemia	V	The benchmark's preoperative anemia label, indicating reduced hemoglobin or red-cell mass.
eGFR < 60	V	Estimated glomerular filtration rate below 60 mL/min/1.73 m^2, indicating reduced kidney filtration.

Clinical prediction targets. The 15 classification and 17 regression targets in Table 22 use recorded categories and numerical measurements, respectively (Tables 16–17). All are visit-level except VitalDB $EtCO_2$ and BIS, which are window-level targets from capnography and EEG monitoring. MC-MED disease and device labels use prior ICD-10 records. Its laboratory targets use the first available ED-visit value; eGFR is calculated with the race-free CKD-EPI 2021 equation from creatinine, age, and sex. For MIMIC-III, laboratory matching takes the earliest available value from the 24 hours preceding the recording and the recording itself; a measurement during the recording can therefore be retained when no prior value is available. VitalDB albumin is preoperative. These tasks evaluate agreement with clinical records and measurements at their specified times, rather than uniformly prospective outcome prediction.

Table 16: **Clinical classification targets.** These rows cover all 15 dataset–task pairs in Table 22; sex occurs in all three datasets. Positive disease labels refer to recorded conditions.

Task	Data	Definition and clinical interpretation
Shared demographic target		
Sex	M, I, V	Recorded patient sex.
MC-MED: ED disposition and prior clinical history		
ED admission	M	Hospital admission following the ED encounter, versus no admission.
Atrial fibrillation	M	Prior atrial fibrillation: disorganized atrial electrical activity associated with an irregular rhythm. The label records history, not necessarily AF in the observed window.
Heart failure	M	Prior heart failure, a clinical syndrome involving impaired cardiac pumping and/or filling.
Cardiac device	M	Prior recorded cardiac-device status.
Chronic kidney disease	M	Prior chronic kidney disease, indicating persistent impairment of kidney structure or function.
Diabetes	M	Prior diabetes mellitus, a disorder of glucose regulation.
Anemia	M	Prior anemia, characterized by reduced hemoglobin or red-cell mass; the target is a diagnosis record, not a threshold applied to current Hb.
MIMIC-III: hospital outcome and diagnostic category		
In-hospital mortality	I	Death during the indexed hospitalization.
Circulatory diagnosis	I	The primary coded hospital diagnosis belongs to the circulatory-system chapter of ICD-9.
VitalDB: preoperative status and postoperative disposition		
ASA ≥ 3	V	Preoperative American Society of Anesthesiologists physical status III or higher; class III denotes severe systemic disease.
Hypertension	V	Recorded preoperative hypertension, indicating persistently elevated blood pressure.
ICU admission	V	Subsequent postoperative admission to intensive care.

Table 17: **Clinical regression targets.** These rows cover all 17 dataset–task pairs in Table 22. Shared measurements are defined once; sampling times and target levels are specified in the accompanying text.

Task	Data	Measurement and clinical interpretation
Demographics and body size		
Age	M, I, V	Patient age in years. MIMIC-III values above 120 are treated as missing because of de-identification shifts.
BMI	V	Body mass index: weight divided by squared height (kg/m^2), a measure of body size relative to height.
Kidney function, hematology, and blood chemistry		
eGFR	M	Estimated glomerular filtration rate ($mL/min/1.73\,m^2$), reflecting kidney filtration function. This target is continuous.
Hemoglobin (Hb)	M, I	Blood concentration of the oxygen-carrying protein in red blood cells.
Albumin	M, V	Serum concentration of albumin, a major protein synthesized by the liver.
BUN	M, I	Blood urea nitrogen: the concentration of nitrogen in circulating urea, related to nitrogen metabolism and renal handling.
Potassium	M	Blood potassium concentration, relevant to electrical activity of nerve and muscle cells, including the heart.
Calcium	M	Blood calcium concentration, relevant to neuromuscular function and multiple cellular processes.
Sodium	M	Blood sodium concentration, relevant to water balance and extracellular fluid regulation.
Glucose	M	Blood glucose concentration, reflecting glycemic status at the sampled time.
Independent intraoperative monitoring channels		
$EtCO_2$	V	End-tidal carbon dioxide partial pressure measured by capnography (mmHg), describing expired CO_2 at the end of a breath.
BIS	V	EEG-derived, unitless bispectral index, reflecting the hypnotic effect of anesthesia.

D.3 Cross-modal retrieval

We extend the 100-candidate evaluation in Table 4 to pools of 2,000 candidates, testing how well PPG-LM identifies paired signals and descriptions as the search space grows. Table 18 reports both signal-to-text (S2T) and text-to-signal (T2S) retrieval on MC-MED, MIMIC-III, and VitalDB. Candidates are ranked by cosine similarity in the aligned embedding space. We report R@1, R@5, and R@10: the fraction of queries whose paired target appears among the top one, five, or ten candidates. The 100-candidate results average two disjoint pools, each with one window per patient and distinct reference captions.

Effect of candidate-pool size. Expanding the pool from 100 to 2,000 candidates reduces recall in both directions, with larger R@1 declines on the external datasets. On MC-MED, S2T/T2S R@1 decreases from 0.975/0.965 to 0.780/0.744, while R@10 remains 0.998/0.995. At the same pool size, MIMIC-III reaches R@10 of 0.853/0.762 and VitalDB reaches 0.510/0.483. Thus, matching the exact pair becomes more difficult as the candidate pool grows, particularly after transfer to the external datasets, while MC-MED retains nearly all paired targets within the top ten results.

Table 18: **Bidirectional retrieval with varying candidate pools.** PPG-LM recall (↑) with $N = 100$ or 2,000 candidates. S2T: signal to text; T2S: text to signal.

Direction	Candidates (N)	R@1↑	R@5↑	R@10↑
MC-MED (ED)				
S2T	100	0.975	1.000	1.000
	2,000	0.780	0.988	0.998
T2S	100	0.965	1.000	1.000
	2,000	0.744	0.980	0.995
MIMIC-III (ICU)				
S2T	100	0.805	1.000	1.000
	2,000	0.258	0.671	0.853
T2S	100	0.755	0.990	1.000
	2,000	0.207	0.560	0.762
VitalDB (OR)				
S2T	100	0.555	0.885	0.965
	2,000	0.108	0.344	0.510
T2S	100	0.550	0.905	0.955
	2,000	0.097	0.324	0.483

D.4 Caption evaluation and numerical read-out

We assess the factual accuracy of PPG descriptions and the accuracy of heart rates read from those descriptions. We compare two PPG-LM interfaces: **direct caption generation**, which produces a new description from the waveform using the caption decoder with greedy decoding; and **nearest-caption retrieval**, which returns the highest-scoring existing caption by embedding similarity. Retrieval uses a fixed bank of 20,000 MC-MED captions from patients disjoint from the evaluation queries, including for external evaluation on MIMIC-III and VitalDB.

The VitalDB analyses here use the official 40 Hz waveform view of the same evaluation items. The main PPG-LM results in Table 3 use the original 500 Hz source, accounting for the different VitalDB values in Tables 19 and 20.

Caption evaluation. The deterministic parser checks stated fields against reference measurements; a numerical HR statement is correct within ± 5 bpm. Fact accuracy is the number of correct stated fields divided by all stated fields, and coverage is the fraction of gradable fields that are stated. Table 19 uses the six common fields across datasets for fact accuracy. The original MC-MED evaluation additionally grades perfusion and morphology, yielding an eight-field accuracy of 0.932 for direct generation and 0.889 for nearest-caption retrieval, versus 0.923 and 0.877 for the common six fields.

The LLM judge, GPT-5.6-luna, receives the reference fact sheet and candidate caption. It classifies fields as correct, incorrect, or absent, counts unsupported claims, and scores factuality and completeness from 1 to 5. Optional charted vitals serve as supporting evidence but do not count toward completeness. Pairwise evaluation uses both presentation orders (400 judgments for 200 windows); generated PPG-LM captions are preferred to GPT-5.6-luna captions at rates of 0.922, 0.855, and 0.777 on MC-MED, MIMIC-III, and VitalDB. As a calibration check, reference captions receive factuality scores of 4.40, 4.80, and 4.85, with 0.86, 0.23, and 0.17 unsupported claims per caption, respectively, indicating that the judge is an imperfect evaluator. Direct generation has higher parser fact accuracy on all three datasets, whereas nearest-caption retrieval receives higher judge factuality scores and fewer unsupported claims (Table 19).

Table 19: **Caption factuality by output method.** Direct generation produces a new caption; nearest-caption retrieval selects an existing caption from the fixed bank. Fact accuracy covers six common fields; coverage includes all gradable fields. Judge factuality is on a 1–5 scale; unsupported claims are counted per caption. VitalDB uses the 40 Hz view. Bold marks the best value within each dataset.

	Rule-based evaluation		LLM-judge evaluation	
Model / caption source	Fact acc.↑	Coverage↑	Factuality↑	Unsupported↓
MC-MED (ED)				
GPT-5.6-luna	0.499	0.99	2.63	6.54
PPG-LM: direct generation	**0.923**	**1.00**	3.64	2.94
PPG-LM: nearest-caption retrieval	0.877	**1.00**	**3.83**	**1.51**
MIMIC-III (ICU)				
GPT-5.6-luna	0.456	0.95	2.83	3.31
PPG-LM: direct generation	**0.920**	**1.00**	3.59	2.88
PPG-LM: nearest-caption retrieval	0.859	**1.00**	**3.67**	**1.47**
VitalDB (OR)				
GPT-5.6-luna	0.486	0.98	2.88	3.21
PPG-LM: direct generation	**0.837**	0.98	3.30	3.17
PPG-LM: nearest-caption retrieval	0.778	**1.00**	**3.43**	**1.52**

Heart-rate read-out. For each PPG-LM interface, we extract the numerical HR stated in its output caption and compare it with the synchronized ECG reference. Table 20 also includes two digital signal processing (DSP) baselines based on the dominant frequency and detected pulse peaks. HR read from directly generated captions has the lowest MAE and the highest fraction of estimates within 5 bpm on all three datasets. On MC-MED, its MAE is 0.27 bpm against the PPG-derived annotation rate but 3.90 bpm against ECG; the independent ECG reference therefore assesses agreement beyond the caption-construction target.

Table 20: **Heart-rate estimation from waveforms or captions.** DSP estimates HR directly from the waveform; PPG-LM reads HR from a generated or retrieved caption. All entries use 300 valid ECG-referenced windows per dataset. MAE is in bpm; "Within 5 bpm" is a fraction. VitalDB uses the 40 Hz view. Bold marks the best value within each dataset.

Method / HR source	MAE↓	Within 5 bpm↑
MC-MED (ED)		
DSP: dominant frequency	6.07	0.857
DSP: pulse peaks	6.36	0.690
PPG-LM: HR from retrieved caption	4.70	0.847
PPG-LM: HR from generated caption	**3.90**	**0.877**
MIMIC-III (ICU)		
DSP: dominant frequency	6.23	0.823
DSP: pulse peaks	6.40	0.680
PPG-LM: HR from retrieved caption	5.36	0.833
PPG-LM: HR from generated caption	**4.27**	**0.860**
VitalDB (OR)		
DSP: dominant frequency	5.53	0.870
DSP: pulse peaks	7.93	0.723
PPG-LM: HR from retrieved caption	4.74	0.867
PPG-LM: HR from generated caption	**4.08**	**0.900**

D.5 Additional Visit-Level Clinical-Fact Recognition

Complementing Section 5.2 and Table 6, Table 21 reports additional clinical-fact recognition on 400 MC-MED visits with eight PPG windows each. PPG-LM averages frozen pretrained fact embed-

dings across windows and scores them against textual fact descriptions. Read-out thresholds are calibrated on held-out MC-MED validation visits, with no calibration on test visits.

Panel (a) covers eight triage-context tasks: high acuity (ESI 1–2), arrival by ambulance versus self-arrival, three heart-rate (HR) classes, low oxygen saturation ($SpO_2 < 95\%$), high respiratory rate (RR $>$ 20 breaths/min), fever (Temp. $\geq 38\,^{\circ}$C), four blood-pressure (BP) classes, and ten chief-complaint (CC) groups. Balanced accuracy accounts for class imbalance in both binary and multi-class tasks.

Panel (b) covers eleven binary recorded conditions: hypertension (HTN), atrial fibrillation (AF), heart failure (HF), coronary artery disease (CAD), type 2 diabetes mellitus (DM), chronic kidney disease (CKD), anemia, obstructive sleep apnea (OSA), asthma, obesity, and stroke. PPG-LM achieves the highest reported balanced accuracy on every task in both panels, reaching 0.785 for HR and 0.729 for fever, with medical-history scores ranging from 0.580 to 0.687. These results measure associations with recorded visit attributes rather than independent clinical diagnoses.

Table 21: **Additional visit-level clinical-fact recognition.** Balanced accuracy (↑) for (a) triage context and (b) medical history on 400 MC-MED visits using eight windows each. PPG-LM read-outs use thresholds calibrated on held-out MC-MED validation visits. Bold marks column maxima.

(a) Triage-context recognition

Model	ESI↑	Arrival↑	HR↑	SpO_2↑	RR↑	Temp.↑	BP↑	CC↑
GPT-5.6-luna	0.501	0.445	0.463	0.497	0.506	0.500	0.269	0.134
GPT-6-luna	0.500	0.498	0.486	0.500	0.500	0.500	0.238	0.128
GPT-6-sol	0.536	0.509	0.736	0.499	0.506	0.512	0.265	0.126
Opus 5.5	0.501	0.504	0.528	0.500	0.500	0.500	0.259	0.100
PulseLM	0.483	0.500	0.656	0.499	0.500	0.498	0.250	0.113
Qwen3-4B (SFT)	0.547	0.500	0.625	0.500	0.500	0.500	0.250	0.100
PPG-LM	**0.618**	**0.607**	**0.785**	**0.650**	**0.692**	**0.729**	**0.442**	**0.202**

(b) Medical-history recognition

Model	HTN↑	AF↑	HF↑	CAD↑	DM↑	CKD↑	Anemia↑	OSA↑	Asthma↑	Obesity↑	Stroke↑
GPT-5.6-luna	0.500	0.502	0.495	0.460	0.483	0.481	0.496	0.531	0.500	0.478	0.512
GPT-6-luna	0.506	0.499	0.494	0.464	0.525	0.510	0.498	0.513	0.499	0.526	0.500
GPT-6-sol	0.498	0.598	0.524	0.495	0.490	0.500	0.500	0.500	0.500	0.522	0.500
Opus 5.5	0.547	0.659	0.546	0.530	0.487	0.494	0.490	0.500	0.500	0.500	0.500
PulseLM	0.504	0.534	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Qwen3-4B (SFT)	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
PPG-LM	**0.583**	**0.676**	**0.687**	**0.656**	**0.586**	**0.682**	**0.638**	**0.599**	**0.609**	**0.580**	**0.593**

D.6 MEDICATION-EVENT EVALUATION

Main medication-class benchmark. Table 5 uses 393 administrations and 393 controls on MC-MED and 139 administrations and 139 controls on VitalDB. Medication names are mapped to the classes used during training. Each view selects up to three nearest windows within 30 minutes before the anchor and five uniformly spaced windows within 60 minutes after it, retaining signed time offsets. Each reported AUROC compares one medication class with no-administration controls; other medication classes are excluded from that comparison's negative group. These are class-specific ranking tasks, rather than a single measure of whether any medication was administered.

Supplementary smaller-cohort evaluation. The earlier language event benchmark contains 150 administrations and 150 controls on MC-MED, with up to 15 administrations per medication class. Its selection windows are 30 minutes before and 60 minutes after the anchor, with up to three pre-event and five post-event observations. VitalDB contains 60 events and 60 controls within ± 2 minutes; MIMIC-III has no event source in this benchmark. Same-visit paired accuracy uses the 56 MC-MED and 19 VitalDB pairs with both views available, assigning half credit to ties. It is not an accuracy over all event/control items. GPT-5.6-luna has tie fractions 0.70 and 0.53, respectively. Its reported AUROC comes from a Boolean answer and equals balanced accuracy, whereas PPG-

LM uses a continuous difference between the highest medication-class score and the no-medication score. On MC-MED, PPG-LM balanced accuracy is 0.647, compared with 0.550 for the LLM; on VitalDB the values are 0.533 and 0.633. The separate validation diagnostic in the training report uses a different item set (AUROC 0.711, paired accuracy 0.687); those values are not directly comparable to the class-versus-control AUROCs in Table 5.

D.7 Full Clinical Prediction Results

This subsection provides all 32 dataset–task pairs evaluated in Section 5.3: 17 on MC-MED, six on MIMIC-III, and nine on VitalDB, comprising 15 classification and 17 regression tasks. Table 22 retains the complete results; Table 7 presents six tasks per dataset. Task definitions are given in Appendix D.2.

Evaluation protocol. Logistic-regression classifiers and ridge-regression models are fitted on frozen representations using each dataset's training patients and evaluated on its test patients. Classification uses AUROC and regression uses R^2; results are reported as mean $\pm$ standard deviation. External-dataset probes therefore assess supervised transfer of the pretrained representations. For visit-level tasks on MC-MED, MIMIC-III, and VitalDB, all foundation-model baselines (including AnyPPG) and both PPG-LM stages use the same four PPG windows per visit, with evenly spaced windows on MC-MED. Baselines and Stage 1 mean-pool the four window features, while Stage 2 applies its learned visit aggregator to those same four windows. The observation budget is therefore matched across models and stages, while the visit read-out differs. VitalDB $EtCO_2$ and BIS are evaluated on the same individual windows across models.

MC-MED diagnosis targets denote recorded prior diagnoses. Its eGFR target is derived from creatinine, age, and sex and is not independent of the demographic targets. Laboratory timing differs across datasets: MC-MED uses the first ED-visit value, MIMIC-III uses the available admission/recording-associated value, and VitalDB uses preoperative measurements. MIMIC-III's matching can retain a value measured during the recording when no earlier value is present. These probes therefore assess association with recorded clinical labels rather than a uniform prospective endpoint.

Full clinical prediction results. Stage 2 exceeds all six foundation-model baselines in mean performance on all 32 tasks and improves over Stage 1 on 30. Stage 1 remains slightly stronger on VitalDB hypertension (0.656 vs. 0.655) and $EtCO_2$ (0.153 vs. 0.152). Despite these relative gains, low R^2 for electrolytes, glucose, and BMI indicates limited predictive performance for these targets.

D.8 Few-Shot Adaptation

We assess how labeled support affects adaptation across MC-MED, MIMIC-III, and VitalDB, complementing the clinical-transfer evaluation in Section 5.3.

Few-shot adaptation of PPG-LM. We fit linear classifiers on frozen representations using $k \in \{5, 10, 20, 50\}$ labeled examples per class from a support pool that is patient-disjoint from the query set. Table 23 reports macro-F1 as mean $\pm$ standard deviation over five support draws; columns correspond to the four per-class label budgets. Tasks cover rhythm regularity, heart-rate (HR) class, sex, and dataset-specific outcomes: ED admission on MC-MED, in-hospital mortality on MIMIC-III, and ICU admission on VitalDB. HR-class performance increases consistently with more support, reaching 0.921, 0.912, and 0.873 at $k = 50$ on the three datasets, respectively. Eleven of twelve tasks attain their highest mean at $k = 50$; the exception is VitalDB ICU admission, which peaks at $k = 10$ (0.591) and falls to 0.557 at $k = 50$. Thus, additional labels generally improve adaptation, but gains remain task-dependent.

For comparison, the separate LLM in-context protocol uses ten labeled windows in total, rather than k examples per class. On MC-MED, PulseLM's linear probes at $k = 10/50$ reach rhythm macro-F1 of 0.395/0.403 and HR-class macro-F1 of 0.750/0.811. These evaluations use distinct adaptation procedures and should be interpreted under their respective support budgets.

Table 22: **Full clinical prediction results.** Test AUROC and R^2 (mean ± SD; ↑) for all 32 tasks. All models use the same four windows for visit-level targets. Bold: best mean; †: window-level target.

	Foundation models						PPG-LM	
Task	PaPaGei-S	PulsePPG	AnyPPG	SIGMA-PPG	MOMENT-L	Chronos-2	Stage 1	Stage 2
MC-MED (ED) — AUROC ↑								
Sex	0.692±0.006	0.719±0.007	0.783±0.008	0.712±0.008	0.724±0.005	0.739±0.008	0.813±0.006	**0.856**±0.005
ED admission	0.673±0.011	0.681±0.009	0.750±0.013	0.679±0.011	0.700±0.011	0.728±0.013	0.765±0.012	**0.787**±0.010
Atrial fibrillation	0.742±0.012	0.752±0.013	0.845±0.010	0.755±0.013	0.768±0.014	0.815±0.012	0.857±0.010	**0.870**±0.011
Heart failure	0.708±0.018	0.723±0.015	0.825±0.013	0.711±0.017	0.731±0.015	0.781±0.012	0.836±0.010	**0.860**±0.009
Cardiac device	0.713±0.023	0.729±0.018	0.835±0.022	0.733±0.022	0.719±0.017	0.784±0.023	0.836±0.022	**0.857**±0.025
Chronic kidney disease	0.647±0.017	0.666±0.016	0.761±0.016	0.666±0.014	0.673±0.013	0.735±0.013	0.774±0.013	**0.792**±0.014
Diabetes	0.644±0.012	0.649±0.014	0.717±0.015	0.660±0.020	0.661±0.016	0.698±0.016	0.728±0.016	**0.740**±0.016
Anemia	0.608±0.008	0.604±0.011	0.668±0.014	0.620±0.010	0.621±0.008	0.654±0.010	0.686±0.015	**0.709**±0.012
MC-MED (ED) — R^2 ↑								
Age	0.395±0.016	0.428±0.018	0.626±0.020	0.389±0.017	0.475±0.017	0.521±0.016	0.691±0.016	**0.782**±0.014
eGFR	0.155±0.015	0.173±0.016	0.320±0.020	0.169±0.012	0.202±0.017	0.262±0.017	0.356±0.021	**0.404**±0.024
Hemoglobin	0.083±0.011	0.098±0.013	0.196±0.013	0.102±0.012	0.123±0.014	0.164±0.014	0.238±0.014	**0.274**±0.012
Albumin	0.093±0.015	0.094±0.011	0.207±0.015	0.097±0.012	0.122±0.016	0.185±0.017	0.239±0.018	**0.286**±0.013
BUN	0.053±0.009	0.062±0.015	0.152±0.015	0.063±0.009	0.070±0.014	0.114±0.010	0.178±0.019	**0.213**±0.019
Potassium	0.025±0.006	0.023±0.009	0.057±0.008	0.029±0.006	0.025±0.008	0.041±0.008	0.067±0.007	**0.086**±0.010
Calcium	0.002±0.006	0.011±0.010	0.041±0.008	0.010±0.005	0.001±0.008	0.024±0.009	0.051±0.009	**0.076**±0.012
Sodium	0.017±0.010	0.023±0.008	0.056±0.010	0.026±0.008	0.017±0.007	0.045±0.010	0.068±0.012	**0.091**±0.015
Glucose	0.032±0.007	0.039±0.007	0.065±0.008	0.041±0.009	0.028±0.007	0.049±0.011	0.071±0.007	**0.083**±0.011
MIMIC-III (ICU) — AUROC ↑								
Sex	0.651±0.024	0.686±0.014	0.713±0.018	0.672±0.016	0.682±0.019	0.664±0.020	0.730±0.019	**0.733**±0.018
In-hospital mortality	0.697±0.029	0.741±0.034	0.756±0.031	0.732±0.025	0.751±0.025	0.772±0.020	0.786±0.022	**0.793**±0.021
Circulatory diagnosis	0.666±0.015	0.690±0.022	0.725±0.017	0.658±0.018	0.672±0.012	0.677±0.017	0.720±0.022	**0.726**±0.023
MIMIC-III (ICU) — R^2 ↑								
Age	0.295±0.039	0.308±0.031	0.468±0.029	0.288±0.036	0.347±0.036	0.367±0.028	0.502±0.030	**0.507**±0.028
BUN	0.029±0.011	0.029±0.009	0.065±0.019	0.034±0.015	0.055±0.015	0.074±0.012	0.081±0.012	**0.086**±0.013
Hemoglobin	0.035±0.010	0.048±0.010	0.079±0.023	0.034±0.009	0.046±0.012	0.076±0.018	0.092±0.024	**0.095**±0.020
VitalDB (OR) — AUROC ↑								
Sex	0.586±0.023	0.626±0.022	0.658±0.016	0.616±0.016	0.620±0.017	0.608±0.016	0.684±0.012	**0.691**±0.013
ASA ≥ 3	0.606±0.021	0.642±0.020	0.690±0.026	0.605±0.035	0.625±0.035	0.655±0.040	0.699±0.034	**0.704**±0.025
Hypertension	0.598±0.016	0.607±0.020	0.635±0.019	0.595±0.025	0.610±0.016	0.617±0.023	**0.656**±0.016	0.655±0.017
ICU admission	0.690±0.009	0.700±0.004	0.747±0.002	0.683±0.009	0.662±0.011	0.714±0.009	0.733±0.004	**0.751**±0.006
VitalDB (OR) — R^2 ↑								
Age	0.240±0.026	0.263±0.036	0.403±0.024	0.259±0.026	0.302±0.025	0.337±0.032	0.418±0.030	**0.435**±0.029
BMI	0.015±0.005	0.022±0.003	0.036±0.008	0.026±0.005	0.035±0.003	0.013±0.010	0.041±0.007	**0.043**±0.007
Albumin	0.075±0.026	0.085±0.034	0.136±0.046	0.085±0.037	0.089±0.038	0.101±0.040	0.134±0.054	**0.138**±0.048
$EtCO_2$†	0.098±0.023	0.121±0.034	0.105±0.028	0.113±0.024	0.133±0.029	0.150±0.031	**0.153**±0.038	0.152±0.037
BIS†	0.380±0.019	0.398±0.022	0.393±0.021	0.387±0.020	0.402±0.021	0.404±0.024	0.407±0.022	**0.408**±0.022

Table 23: **Few-shot adaptation of PPG-LM.** Macro-F1 (mean ± SD; ↑) over five support draws. VitalDB uses the 40 Hz waveform view. k: examples per class. Bold: best mean per row.

Task	$k = 5$	$k = 10$	$k = 20$	$k = 50$
MC-MED (ED) — Macro-F1 ↑				
Rhythm	0.488±0.051	0.537±0.019	0.567±0.037	**0.597**±0.022
HR class	0.755±0.023	0.831±0.025	0.898±0.018	**0.921**±0.003
Sex	0.568±0.042	0.645±0.054	0.696±0.041	**0.710**±0.026
ED admission	0.563±0.014	0.558±0.026	0.583±0.018	**0.617**±0.008
MIMIC-III (ICU) — Macro-F1 ↑				
Rhythm	0.506±0.029	0.585±0.037	0.572±0.025	**0.591**±0.021
HR class	0.757±0.050	0.859±0.023	0.891±0.017	**0.912**±0.006
Sex	0.505±0.058	0.538±0.039	0.562±0.040	**0.580**±0.033
In-hospital mortality	0.548±0.053	0.576±0.042	0.597±0.042	**0.636**±0.015
VitalDB (OR) — Macro-F1 ↑				
Rhythm	0.375±0.023	0.365±0.013	0.360±0.020	**0.399**±0.006
HR class	0.761±0.039	0.832±0.043	0.862±0.026	**0.873**±0.011
Sex	0.498±0.018	0.501±0.039	0.496±0.035	**0.558**±0.013
ICU admission	0.543±0.033	**0.591**±0.017	0.552±0.045	0.557±0.041

D.9 OBJECTIVE ABLATIONS

Table 24: **Objective ablations on MC-MED.** Rhythm/quality: macro-F1; retrieval: signal-to-text R@1 with 2,000 candidates. Stage-2 CLIP/CoCa report per-metric bests across configurations.

Variant	Rhythm ↑	Quality ↑	R@1 ↑
Stage 1			
CLIP	0.650	0.505	0.790
SigLIP	0.620	0.400	0.815
CoCa	0.640	0.614	0.783
Stage 2			
CLIP	0.534	0.580	0.797
SigLIP	0.600	0.470	**0.820**
CoCa	**0.651**	**0.673**	**0.820**

Table 24 compares three segment objectives across training stages. CLIP (Radford et al., 2021) uses symmetric InfoNCE, $\mathcal{L}_{\text{CLIP}} = \mathcal{L}_{\text{NCE}}(Q; \mathcal{C})$, while SigLIP (Zhai et al., 2023) uses the pairwise sigmoid loss $\ell_{ij} = -\log \sigma[y_{ij}(Q_{ij} + b)]$, where $y_{ij} \in \{+1, -1\}$ labels matched/mismatched pairs and b is a learned bias. CoCa additionally optimizes caption likelihood. At Stage 1, CLIP leads rhythm recognition, SigLIP leads retrieval, and CoCa leads quality assessment. At Stage 2, CoCa improves all three metrics, achieving the highest rhythm and quality macro-F1 and matching SigLIP's R@1. This combination supports CoCa for attribute recognition and retrieval while retaining caption generation. The Stage-2 CoCa R@1 of 0.820 uses signal-only captions; the main model's signal-plus-vitals configuration reaches 0.780 at the same candidate-pool size (Table 18).

Component-ablation setting. Table 8 uses the CLIP configuration on MC-MED validation, with shared Stage-1 initialization and matched Stage-2 training budgets. Temporal accuracy distinguishes true from corrupted statements; event AUROC compares any medication event against no-administration controls; segment R@1 measures segment–text retrieval.

D.10 PROMPT SENSITIVITY

We examine whether expanding the class descriptions improves zero-shot recognition while keeping the model and evaluation items fixed. The *Template* bank contains the caption-style descriptions used in the main experiments; the *Combined* bank augments them with reworded variants. For each class, we average its text embeddings and score the resulting representation against the waveform embedding by cosine similarity. Table 25 reports macro-F1 for three-class rhythm, five-class heart rate (HR), and three-class signal quality on MC-MED, MIMIC-III, and VitalDB.

Template prompts achieve higher rhythm and quality scores on all three datasets. Combined prompts improve HR-class macro-F1 from 0.355 to 0.393 on MC-MED and from 0.337 to 0.350 on MIMIC-III, but reduce it from 0.327 to 0.282 on VitalDB. The template bank thus performs better on seven of the nine dataset–task pairs. Expanding the prompt bank does not consistently improve recognition, indicating that the choice of class descriptions remains consequential even when the model is unchanged.

Table 25: **Prompt sensitivity of PPG-LM.** Template uses caption-style class descriptions; Combined adds reworded variants. Macro-F1 (↑); bold marks the best score per column.

	MC-MED			**MIMIC-III**			**VitalDB**		
Prompt bank	Rhythm	HR class	Quality	Rhythm	HR class	Quality	Rhythm	HR class	Quality
Template	**0.651**	0.355	**0.673**	**0.670**	0.337	**0.786**	**0.389**	**0.327**	**0.583**
Combined	0.629	**0.393**	0.592	0.642	**0.350**	0.677	0.384	0.282	0.515

E LANGUAGE TEMPLATES AND EXAMPLES

This appendix presents the language used for training and evaluation. Training captions are rendered from structured measurements and clinical records using fixed template banks; the example identifier and random seed determine template selection and sentence order. No external LLM is used. We distinguish these training targets from zero-shot class descriptions and the model's learned text-prefix embeddings.

E.1 REPRESENTATIVE CAPTION TEMPLATES

Caption content. Table 26 reproduces representative template strings, with braces marking substituted fields. Segment captions describe available rate, rhythm, and quality evidence; missing features are omitted, and poor-quality segments receive a quality-only description. Event captions identify the medication class associated with the sampled windows. Visit captions combine available demographics, triage context, and prior history.

Table 26: **Representative training templates.** Strings are reproduced verbatim; braced fields are filled from measurements or records.

Target	Template
Segment rate	The pulse rate is about {val} beats per minute, {level}.
Segment rhythm	The cardiac rhythm is {level}.
Segment quality	The PPG trace is {level}.
Poor-quality segment	This segment's pulse waveform is too noisy to characterize reliably.
Medication event	During this period the patient was given {desc}.
Control interval	This period of monitoring contains no medication administration.
Chief complaint	The chief complaint at triage was {cc}.

Temporal captions and negatives. Temporal templates express occurrence, prevalence, contiguous duration, or temporal order. A negative caption replaces one eligible sentence and retains the others. For example, the implemented positive sentence "Over the stay the pulse rate trended upward." can be replaced by "Over the stay the pulse rate trended downward." This pair illustrates the corruption rule, rather than an observed trajectory.

E.2 ZERO-SHOT EVALUATION PROMPTS

Segment scoring. Table 27 reproduces representative prompts from the language benchmark in Section 5. The main segment-recognition results use two caption-style descriptions per class, whose embeddings are averaged for cosine scoring. Appendix D.10 compares this template bank with a combined bank that adds reworded class descriptions. Quality recognition uses a separate clean/fair/noisy prompt bank.

Visit and event scoring. Visit prompts describe individual attributes and are scored against the selected visit read-out. The MC-MED event bank contains ten medication classes and a no-medication description. For the supplementary any-administration diagnostic in Appendix D.6, the score is the highest medication-class score minus the no-medication score. The main medication table instead reports separate class-versus-control AUROCs.

Table 27: **Representative zero-shot prompts.** Verbatim examples from the segment template bank and the visit and event class descriptions, grouped by task.

Concept	Prompt
Pulse rate	
Marked bradycardia	The measured pulse rate is markedly bradycardic.
Bradycardia	The measured pulse rate is bradycardic.
Normal rate	The measured pulse rate is within the normal range.
Mild tachycardia	The measured pulse rate is mildly tachycardic.
Tachycardia	The measured pulse rate is tachycardic.
Rhythm	
Regular rhythm	The pulse rhythm appears regular.
Irregular rhythm	The pulse rhythm appears slightly irregular.
Marked irregularity	The pulse rhythm appears markedly irregular.
Signal quality	
Clean signal	Pulse-wave quality here is clean and reliable.
Fair signal	Pulse-wave quality here is of fair quality.
Noisy signal	The signal is somewhat noisy but usable for analysis.
Visit attributes	
Visit acuity	The visit is an emergent, high-acuity presentation.
Visit demographics	The recording is from a male patient.
Visit complaint	The visit is for a patient presenting with chest pain.
Medication events	
Medication class	The patient received a bronchodilator during this period.
Control	No medication was given during this period.

E.3 Synthetic Caption Examples and Decoder Context

These captions were rendered from invented feature values and clinical fields using the captioning pipeline. They illustrate training targets; they are neither patient records nor model-generated outputs. Missing fields contribute no sentences.

Segment caption.

> Overall the PPG signal is clean and reliable. Over the window the pulse rate stays stable. The heartbeat is regular across the window. Heart rate estimated from the pulse waveform is around 82 bpm, within the normal range.

Event caption.

> This segment covers the time around a dose of an inhaled bronchodilator.

Visit caption with history.

> The individual is a 67-year-old male patient (older adult). Arrival vital signs: heart rate 108 bpm (mildly tachycardic), SpO2 96% (normal). The patient got to the ED by ambulance (EMS). Emergency severity was emergent, ESI level 2. Chief complaint: shortness of breath. Past medical history includes essential hypertension. The home medication list includes statins.

CoCa decoder context. CoCa tokenizes each segment caption into a beginning-of-sequence token, caption subwords, and an end-of-sequence token. The shared text encoder prepends the learned segment prefix e_{seg} (denoted `[WIN]` in the implementation), which is a vector rather than a literal word. After removing the prefix position, the causal text states enter the separate multimodal decoder, which cross-attends to waveform memory M_i and predicts the next caption token; padding targets are ignored. Training uses no natural-language generation prompt. The language benchmark additionally evaluates greedy waveform-conditioned decoding alongside retrieval and class-prompt scoring.