SENSOR-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sensors are useful not only for understanding the world but also for deciding what to do *next*. Existing sensor models however largely stop at *perception*: they recognize states or predict outcomes, leaving *actions* modeled separately through task-specific and often closed label spaces. We introduce **Sensor-Language-Action (SLA)** modeling, a framework that connects multi-modal sensor observations, natural language, and actions within a shared model. SLA uses language as a semantic interface between sensing and acting, allowing heterogeneous actions to be represented, predicted, and explained while remaining grounded in the underlying sensor evidence. We build a large-scale SLA benchmark spanning more than 116,000 individuals, 79 sensor modalities, and 60 action groups, together with a multi-faceted captioning pipeline that aligns individual context, sensor dynamics, and action evidence. Building on this framework, we present `OpenSLA`, a unified SLA model for hierarchical action prediction, state understanding, and action explanation. Extensive experiments on real-world tasks in clinical prediction, operating rooms, and metabolic health verify its superior performance over state-of-the-arts. `OpenSLA` also demonstrates intriguing capabilities including language-guided evidence grounding and zero-shot generalization to unseen actions and cohorts.

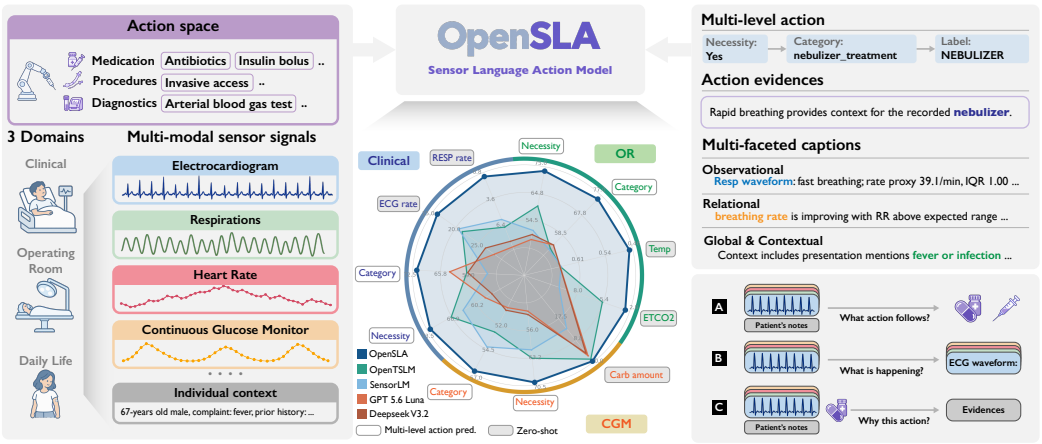

Figure 1: **Sensor-Language-Action (SLA) Model.** We establish SLA as a unified paradigm for sensor-action modeling through a language interface over sensor observations, patient context, and actions. We instantiate SLA with `OpenSLA` and evaluate it across more than 116,000 individuals, 79 sensor modalities, and 60 action groups. In addition to the superior performance over state-of-the-art methods, `OpenSLA` supports **(A)** multi-level action predictions, **(B)** sensor-state understanding, and **(C)** action-evidence grounding.

1 INTRODUCTION

Intelligent systems do more than *perceive*: they observe, understand, and ***act***. The value of sensing, likewise, lies not only in describing what is happening, but in informing the actions that follow (Xu et al., 2026a; Khasentino et al., 2025). For example, clinicians continuously integrate physiological signals with individual context, history, and prior interventions to decide whether to administer a medication, escalation, or continue observation (Johnson et al., 2022).

Yet, existing sensor models are centered on observing and understanding. Early methods map sensor signals to fixed labels or outcomes (Narayanswamy et al., 2025; Shuai et al., 2026), while recent

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.

sensor-language models enable richer semantic descriptions and open-ended reasoning over sensor observations (Zhang et al., 2025; Langer et al., 2026; Xu et al., 2026b). However, these advances largely stop before **action**: downstream decisions are still modeled as separate prediction problems, each with its own formulation and action space (Lee et al., 2022b; Weidman et al., 2025). As a result, existing approaches provide no common mechanism for relating heterogeneous sensor evidence, contextual information, and the diverse actions they may support.

Natural language, on the other hand, offers a natural bridge between sensing and action. Unlike fixed labels, language can jointly express task intent, individual context, physiological state, and action semantics, allowing heterogeneous sensor-action problems to be represented in a common form. This suggests a broader modeling paradigm in which sensor observations are grounded in *language* and connected directly to *actions*, rather than treating understanding and decision-making as separate stages. We therefore ask:

Can sensor data, language, and action be modeled through a common framework?

To fill the gap, we introduce **Sensor-Language-Action (SLA)** modeling, a unified framework that connects sensor observations, natural-language context, and actions through a common multi-modal interface (see Fig. 1). We further instantiate SLA with `OpenSLA`, to our knowledge, the first general framework for jointly modeling sensor observations, language, and actions. `OpenSLA` extends sensor intelligence from understanding observations to reasoning about and explaining the actions they support, enabled by three key contributions: ❶ A hierarchical, multi-faceted physiological captioning pipeline that aligns local sensor dynamics, cross-signal relations, individual context, and action-relevant evidence as language supervision; ❷ The curation of a large-scale benchmark spanning more than 116,000 individuals, 79 sensor modalities, 3 healthcare domains, 5 tasks, and 60 action groups; and ❸ a unified SLA architecture for hierarchical action prediction, state understanding, and action explanation that supports robust and scalable sensor-action modeling via language grounding.

To rigorously evaluate `OpenSLA`, we benchmark it against state-of-the-art methods across six real-world cohorts and five core tasks spanning clinical care, operating rooms, and daily-life metabolic health. Extensive experiments demonstrate consistently strong performance in hierarchical action prediction, state understanding, and action explanation. Beyond task performance, further analyses reveal that `OpenSLA` can ground actions in relevant sensor evidence and generalize zero-shot to unseen actions and cohorts. Our contributions are as follows:

- We formulate Sensor-Language-Action (SLA) modeling, a new framework that unifies sensor-action problems through multi-modal language prompting, and construct a large-scale benchmark spanning more than 116,000 individuals, 79 sensor modalities, 5 tasks, and 60 action groups.
- We introduce a multi-faceted physiological captioning pipeline that jointly captures local sensor dynamics, cross-signal relations, individual-level context, and action-relevant evidence, providing language supervision tailored to sensor-action modeling.
- We develop `OpenSLA`, a unified SLA model that jointly supports hierarchical action prediction, state understanding, and action explanation, informed by systematic studies of how sensor signals, language, and actions should be modeled together.
- We conduct extensive evaluations across six real-world cohorts, demonstrating state-of-the-art performance across all five core tasks, together with strong sensor-evidence grounding and zero-shot generalization to unseen actions and cohorts.

2 Related Work

Sensor-Action Modeling. Predicting subsequent actions or decisions from observed sensor signals has been a long-standing problem (Lee et al., 2022b; Weidman et al., 2025). Existing approaches typically formulate this as supervised prediction, either learning direct mappings from sensor observations to action labels or incorporating additional textual context as conditioning information (Suresh et al., 2017; Hardan et al., 2024). However, these approaches are less effective and flexible in incorporating heterogeneous individual context and other decision-relevant information. In contrast, we address this limitation by introducing *language* as a flexible interface for sensor-action modeling.

Sensor and Time-Series Foundation Models. Time-series foundation models (Ansari et al., 2025; Woo et al., 2024; Goswami et al., 2024) have shown strong performance across diverse sensor-

Table 1: **Comparisons with representative sensor-language studies.**

Study	Domain			Data Scale				Caption Content	
	Clinical	OR	Daily	# Dataset	# Modality	# Individuals	# Action	Individual Context	Action Evidence
Khasentino et al. (2025)	✗	✗	✓	3	20	4.2k	–	✓	✗
Langer et al. (2026)	✓	✗	✓	5	3	N/R	–	✓	✗
Li et al. (2025b)	✗	✗	✓	5	3	0.2k	–	✗	✗
Zhang et al. (2025)	✗	✗	✓	4	5	103k	–	✗	✗
Xu et al. (2026b)	✓	✗	✗	5	12	10k	–	✗	✗
`OpenSLA` **(Ours)**	✓	✓	✓	**7**	**79**	**116k**	**60**	✓	✓

centered tasks, including forecasting, classification, and representation learning. Beyond general-purpose time-series models, sensor-focused foundation models further specialize in physiological and wearable signals, covering ECG (Li et al., 2025a) and PPG (Pillai et al., 2025), continuous glucose monitoring (Lutsker et al., 2026), motion and wearable sensing (Xu et al., 2026a), and sleep (Shuai et al., 2026). Despite this progress, existing sensor foundation models largely focus on understanding states of input sensor signals, rather than translating sensor dynamics into action-relevant semantics. In contrast, `OpenSLA` addresses this gap by jointly modeling sensor observations, language context, and actions through a shared sensor-language-action interface.

Sensor-Language Modeling. Language models have increasingly been applied to sensor analysis through direct prompting or dedicated sensor-language models (Li et al., 2025b; 2026; Langer et al., 2026). SensorLM (Zhang et al., 2025) and SleepLM (Xu et al., 2026b) leverage paired sensor-text data and multi-level captioning to align sensor representations with language, demonstrating strong transferability across downstream tasks. OpenTSLM (Langer et al., 2026), in contrast, combines time-series encoders with pretrained LLMs to support zero-shot language-based interaction with medical signals. Such models often rely on captioning pipelines for paired sensor-text supervision, but both their data and modeling pipelines are primarily designed for sensor-text alignment rather than action-relevant semantics. In this project, we introduce a multi-faceted captioning pipeline to capture a broader spectrum of information, forming the basis of our sensor-language-action model.

3 Sensor-Language-Action Dataset Construction

Data Regime. We construct our sensor-action testbed from six development datasets and one external dataset spanning three healthcare settings: clinical care, operating room (OR), and continuous glucose monitoring (CGM). The clinical setting includes MC-MED (Kansal et al., 2025), MIMIC-III (Johnson et al., 2016), and MIMIC-IV (Johnson et al., 2022); the OR setting includes MOVER (Samad et al., 2023) and VitalDB (Lee et al., 2022a); and the CGM setting includes MetaboNet (Wolff et al., 2026) and PEDAP (Wadwa et al., 2023). MIMIC-IV (Johnson et al., 2022) is held out as an external clinical cohort for evaluating cross-cohort generalization. For the six development datasets, we construct individual-disjoint train, validation, and test splits and combine the training data within each setting. The clinical and OR domains contain heterogeneous waveform and numerical measurements, with representative signals including ECG, plethysmography, respiration, blood pressure, SpO_2, and heart rate. The CGM domain includes glucose time series together with basal-insulin delivery.

Unified Multi-level Action Space. We define an action as a recorded healthcare decision or intervention at a given decision point. Across the clinical and OR domains, actions span medications, procedures, diagnostic evaluations, and treatment adjustments, while CGM actions correspond to insulin bolus delivery. We organize these actions across datasets into a shared three-level hierarchy: ❶ **action necessity**, indicating whether an action is recorded; ❷ **action category**, grouping related actions with shared clinical or functional semantics; and ❸ **fine-grained action label**, referring to the specific action within a category. For example, an Imipenem administration can be represented as *Yes* at the necessity level, *antibiotics* at the category level, and *Imipenem* at the fine-grained level. Based on this hierarchy, we formulate action prediction from coarse to fine, asking whether an action should be taken, which type of action to take, and which specific action to take within that category. Detailed dataset statistics and action definitions are provided in Appendix A.

Multi-faceted Captioning Pipeline. We aim to capture physiological evidence from sensor signals together with individual context relevant to clinical decisions. Specifically, we organize each generated caption into four complementary facets, comprising sensor observations, physiological dynamics, individual-level context, and action-relevant evidence, as detailed in Fig. 2. Implementation details are in Appendix B.

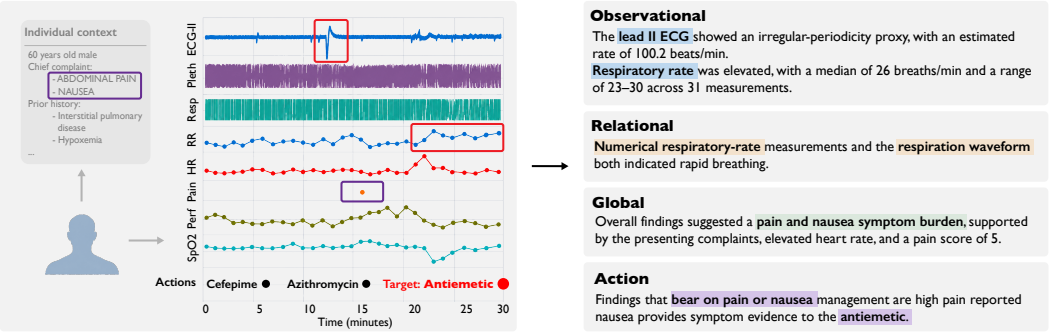

Figure 2: **Multi-faceted sensor-action captioning pipeline.** Our captioning pipeline integrates four complementary facets of sensor-action information: (1) Observational captions summarize signal-derived features; (2) Relational captions describe cross-channel evidence; (3) Global captions capture patient-specific context; and (4) Action captions identify evidence from the preceding facets that is relevant to the action label.

Observational Captions. These captions are constructed from channel-level physiological findings derived using clinically established measurements and waveform characteristics, such as heart rate, respiratory rate, and rhythm patterns. They translate directly observable sensor properties into explicit physiological semantics.

Relational Captions. We further organize physiologically related measurements into complementary feature sets and describe their temporal and cross-channel relationships. For example, respiratory rate and breathing waveforms are jointly considered to characterize respiratory status, allowing multiple measurements to provide convergent evidence about the same physiological process.

Global Captions. We integrate sensor-derived findings with individual-specific context, including available background information and prior clinical history, to summarize the overall physiological state. This places local sensor evidence within the broader clinical circumstances surrounding a decision.

Action Captions. We use recorded action categories to identify relevant evidence from sensor measurements and individual context, and verbalize how this evidence relates to the corresponding action. This action-conditioned evidence selection explicitly connects physiological state with decision-relevant information.

Taken together, as summarized in Table 1, our study covers broader data settings and a larger cohort than the compared sensor-language studies. Across the three domains, our testbed spans six development datasets and the external MIMIC-IV cohort, 79 sensor modalities, 116k individuals, and 60 unique actions, providing broader data coverage and richer information than representative sensor-language studies.

4 On Unifying Sensor-Action Tasks with Language Interface

Sensor-Language-Action Modeling. We begin with a fundamental question: *How can we unlock the world knowledge encoded in pretrained language models for sensor-action modeling?* We therefore compare four large language model (LLM)-centric modeling paradigms: ❶ Zero-shot inference with multi-modal LLMs;❷ VLA-style (Kim et al., 2024) fine-tuning with action supervision only; ❸ SensorLM (Zhang et al., 2025) style signal-language alignment; ❹ Fine-tuning with joint action-and-language supervision. We evaluate these approaches on action prediction using VitalDB, and found action-and-language fine-tuning shows consistent advantages across both comparisons. It substantially outperforms action-only fine-tuning (Fig. 7) and maintains a substantial margin over zero-shot multi-modal LLMs, including GPT-5.6-Luna and DeepSeek-V3.2 (Table 3). Motivated by these observations, we establish the **Sensor-Language-Action (SLA)** paradigm, which jointly models sensor observations, natural-language instructions, and individual context, with both action and language as prediction targets.

We further investigate which sensor-language alignment strategy best supports SLA modeling. Specifically, we compare three representative LLM adaptation approaches on VitalDB action prediction: ❶ LLaVA-style adapter training (Liu et al., 2023); ❷ Flamingo-style cross-attention injection (Alayrac et al., 2022); and ❸ direct Low-Rank Adaptation (LoRA) (Hu et al., 2021). As

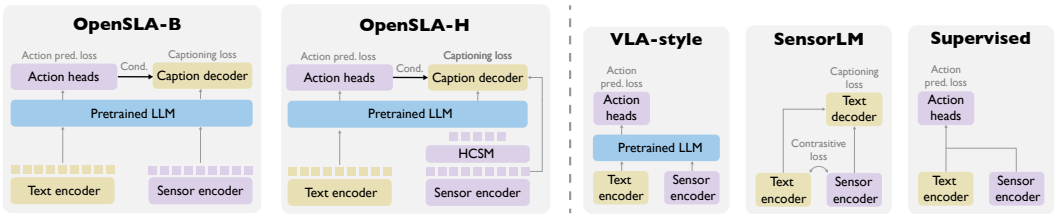

Figure 3: **Comparison of modeling paradigms for sensor-action modeling through language.** `OpenSLA` leverages language knowledge from the pretrained LLM, uses caption supervision to extract action-relevant sensor dynamics, and employs action-conditioned caption decoding to improve action-language alignment.

shown in Fig. 7, direct LoRA adaptation achieves the best performance on most tasks across the OR datasets. Based on these findings, we instantiate `OpenSLA` with the following design.

`OpenSLA`: Learning Sensor Dynamics through a Language Interface. `OpenSLA` consists of three key components: ❶ an LLM backbone equipped with action-prediction heads; ❷ a hierarchical signal encoder; and ❸ a multi-modal fusion decoder. The model is jointly optimized with action prediction and autoregressive language generation objectives. The overall framework is in Fig. 3.

Hierarchical sensor encoder. Sensor recordings often span long durations and vary in channel composition. We propose a hierarchical sensor encoder in which each channel is first encoded independently with a frozen signal encoder, yielding tokens $\mathbf{H}^{(0)} \in \mathbb{R}^{N_0 \times D}$, where N_0 is the token count and D is the encoder feature width. We then introduce a Hierarchical Channel State Memory (HCSM) layer to hierarchically compress these tokens while preserving both sample-level semantics and fine-grained sensor patterns. The first pathway recursively aggregates tokens across multiple resolutions. Given patch-level tokens $\mathbf{H}^{(0)}$, the l-th layer uses N_l learned queries to cross-attend to $\mathbf{H}^{(l-1)}$, producing $\mathbf{H}^{(l)} \in \mathbb{R}^{N_l \times D}$. In parallel, global queries attend directly to the full patch sequence to capture sample-level semantics. Outputs from both paths are concatenated as sensor tokens for the language backbone.

Multi-modal fusion decoder. The hierarchical sensor encoder provides compact multi-resolution representations, but fine-grained sensor details may be attenuated after compression. We therefore introduce a gated residual caption decoder that refines the LLM caption states through two parallel cross-attention branches, over the predicted action profile and retrieved sensor evidence. Each branch is modulated by a learnable gate, and the fused representation is passed to the original language-model head for caption prediction.

Training objective. We denote the action prediction loss as $\mathcal{L}_{\text{act}}$, aggregating losses for action necessity, category, and fine-grained label prediction, and optimize caption generation with autoregressive cross-entropy $\mathcal{L}_{\text{text}}$. The overall objective is $\mathcal{L} = \mathcal{L}_{\text{act}} + \lambda_{\text{text}}\mathcal{L}_{\text{text}}$, where λ_{text} controls the caption-generation loss. During training, we adapt the LLM backbone with LoRA and jointly optimize the hierarchical adapter and multi-modal fusion decoder, while keeping the remaining pretrained modules frozen. Implementation details are provided in Appendix C.1.

5 Experiment & Results

Baselines. We compare `OpenSLA` with four families of sensor modeling methods. ❶ **Supervised models.** We train ConvTrans (Foumani et al., 2024) and MMIM (Han et al., 2021; Hardan et al., 2024) from scratch. ❷ **Time-series foundation models.** We linearly probe Chronos-2 (Ansari et al., 2025) and Moirai (Woo et al., 2024) for action prediction. ❸ **Sensor-language models.** We adapt OpenTSLM (Langer et al., 2026) and SensorLM (Zhang et al., 2025) on our training corpus. ❹ **Multi-modal LLMs.** We prompt GPT-5.6-Luna (OpenAI, 2026) and DeepSeek-V3.2 (DeepSeek-AI et al., 2025) through CodeAct-style Python execution (Wang et al., 2024). Implementation details are provided in the Appendix C.2.

Implementation details. We denote the base and hierarchical variants by `OpenSLA-B` and `OpenSLA-H`, respectively. We adapt a Qwen-3.5-2B backbone with LoRA for all `OpenSLA` variants. For fair comparison, all methods trained on our cohorts use the same training data and compute budget, with signal encoders initialized from the same domain-specific DINO-style pretrained model. Training is limited to 24 hours on H200 GPU nodes. Detailed configurations are in the appendix.

Evaluation. We evaluate OpenSLA against baselines across clinical, operating room (OR), and continuous glucose monitoring (CGM) domains on the following tasks. ❶ **Action Prediction.** We report macro-AUROC and balanced accuracy for action necessity and category prediction, and balanced accuracy for fine-grained action labels. Metrics are macro-averaged across labels within each action granularity level. ❷ **State Understanding.** Following (Xu et al., 2026b), we report MAE between physiological measurements extracted from generated captions and the corresponding ground-truth values. ❸ **Evidence Grounding.** We compare generated reports with the action-evidence portion of the ground-truth captions using BERTScore (Zhang et al., 2019), following prior sensor-language evaluation (Zhang et al., 2025; Xu et al., 2026b). Scores are computed with DistilBERT (Sanh et al., 2019) using the setup of (Xu et al., 2025).

Table 2: **Multi-level action prediction results on clinical domain.**

Method	MC-MED					MIMIC-III				
	Necessity		Category		Fine-label	Necessity		Category		Fine-label
	AUROC↑	BAcc↑	AUROC↑	BAcc↑	BAcc↑	AUROC↑	BAcc↑	AUROC↑	BAcc↑	BAcc↑
Supervised:										
ConvTrans	65.2	57.5	59.8	53.8	54.7	61.4	59.7	54.8	51.7	57.1
MMIM	61.9	53.4	64.6	56.0	53.5	66.0	61.6	64.5	54.4	60.2
TS Foundation Models:										
Chronos-2	62.5	55.4	56.9	52.7	51.8	60.9	56.6	58.8	52.5	53.2
Moirai	62.5	57.4	54.4	51.9	51.7	57.7	55.7	58.7	52.8	53.4
Multi-modal LLMs:										
GPT-5.6-Luna	–	60.6	–	63.1	59.5	–	54.8	–	64.0	**79.5**
DeepSeek-V3.2	–	56.1	–	58.2	54.4	–	55.6	–	60.2	72.2
TS-LLM:										
OpenTSLM-SP	–	62.9	–	52.8	50.1	–	71.2	–	66.0	67.1
SensorLM	–	64.6	–	53.5	50.4	–	62.9	–	55.3	51.7
Ours:										
OpenSLA-H	77.0	72.2	**83.7**	**68.7**	**65.2**	81.9	73.3	88.1	**74.5**	76.5
OpenSLA-B	**77.6**	**72.5**	**83.7**	68.6	61.2	**84.0**	**75.4**	**88.3**	73.1	73.7

Table 3: **Multi-level action prediction results on the operating room domain.**

Method	MOVER					VitalDB				
	Necessity		Category		Fine-label	Necessity		Category		Fine-label
	AUROC↑	BAcc↑	AUROC↑	BAcc↑	BAcc↑	AUROC↑	BAcc↑	AUROC↑	BAcc↑	BAcc↑
Supervised:										
ConvTrans	47.6	48.0	62.0	56.9	56.0	52.1	51.4	70.9	66.0	51.4
MMIM	53.5	51.8	63.2	59.0	55.9	51.7	51.1	63.9	56.8	57.2
TS Foundation Models:										
Chronos-2	43.6	46.3	63.3	60.5	57.5	53.9	52.1	65.7	62.3	58.0
Moirai	46.3	45.7	63.7	57.0	51.9	49.8	49.7	60.9	57.5	54.8
Multi-modal LLMs:										
GPT-5.6-Luna	–	45.0	–	52.4	**69.2**	–	50.1	–	55.7	78.6
DeepSeek-V3.2	–	48.1	–	54.3	55.9	–	50.8	–	53.8	66.2
TS-LLM:										
OpenTSLM-SP	–	71.5	–	55.2	52.3	–	48.9	–	49.9	51.2
SensorLM	–	49.9	–	52.9	50.1	–	52.1	–	51.0	51.7
Ours:										
OpenSLA-H	**85.9**	**80.6**	79.9	**69.9**	67.6	**77.1**	**66.3**	**86.3**	**80.8**	78.5
OpenSLA-B	83.4	74.2	**80.4**	69.0	66.4	74.9	59.0	85.9	79.5	**78.9**

OpenSLA predicts actions across domains and granularities. We evaluate OpenSLA across action necessity, category, and fine-grained action label prediction where applicable. As shown in Table 2, Table 3, and Table 4, OpenSLA consistently outperforms all baselines across all action granularities and datasets. Beyond the overall performance, the comparisons reveal clear domain-specific patterns. Sensor-only action prediction is most challenging in the OR domain, while OpenSLA achieves its largest gains in the Clinical and OR settings, suggesting that textual individual context is particularly valuable for context-dependent clinical actions. Across most settings, prediction becomes easier from action necessity to action category and fine-grained labels, consistent with the increasingly constrained decision space once an intervention is deemed necessary. CGM is a notable exception, where action-category prediction remains challenging because insulin decisions can depend on future carbohydrate intake that is not available in the observed context.

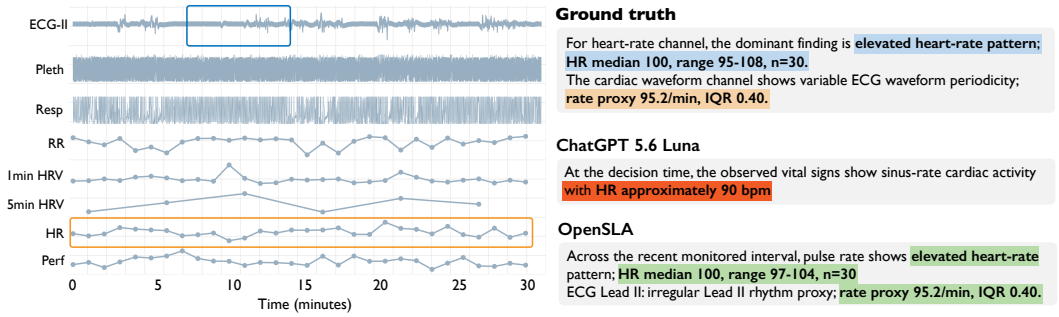

Figure 4: **Sensor-grounded caption generation.** OpenSLA captures both physiological states and fine-grained signal statistics, while the LLM baseline misses key physiological details in the illustrated example.

Table 4: **Multi-level action prediction results on the CGM domain.**

Method	MetaboNet				PEDAP			
	Necessity		Category		Necessity		Category	
	AUROC↑	BAcc↑	AUROC↑	BAcc↑	AUROC↑	BAcc↑	AUROC↑	BAcc↑
Supervised:								
ConvTrans	66.5	62.2	53.8	52.7	79.0	72.8	**70.5**	**66.2**
MMIM	66.3	60.2	62.7	58.7	64.7	61.7	58.5	56.3
TS Foundation Models:								
Chronos-2	63.6	60.3	52.5	51.7	71.7	66.1	55.8	54.0
GluFormer	52.0	51.3	51.0	50.5	53.5	52.2	53.3	52.8
Multi-modal LLMs:								
GPT-5.6-Luna	–	49.9	–	50.4	–	52.0	–	50.1
DeepSeek-V3.2	–	51.2	–	50.5	–	52.5	–	50.6
TS-LLM:								
OpenTSLM-SP	–	62.1	–	53.9	–	64.6	–	51.6
SensorLM	–	62.9	–	56.8	–	59.4	–	51.8
Ours:								
OpenSLA-H	71.9	64.3	**67.2**	**59.6**	80.9	75.1	59.7	53.9
OpenSLA-B	**73.3**	**65.0**	66.5	59.2	**81.2**	**75.5**	65.2	56.3

Table 5: **Zero-shot sensor-state prediction results.** Results are evaluated using MAE↓ across sensor-state prediction tasks.

Method	Clinical			Operating Room			CGM	
	Cardiac Index	ECG-II Rate	RESP Rate	Resp. Rate	Temperature	$ETCO_2$	Basal Rate	Carb Amount
Multi-modal LLMs:								
GPT-5.6-Luna	–	28.42	–	–	–	–	**0.07**	2.7
DeepSeek-V3.2	–	26.48	8.27	–	–	–	0.43	3.7
TS-LLM:								
OpenTSLM-SP	1.13	21.37	6.71	2.66	0.66	5.47	0.58	1.2
SensorLM	–	20.52	5.87	2.56	0.66	9.76	0.54	13.5
Ours:								
OpenSLA-H	1.01	**15.88**	**1.17**	**1.41**	**0.47**	**3.18**	0.35	**0.38**
OpenSLA-B	**0.93**	20.11	1.59	2.48	0.64	5.13	**0.07**	0.73

Language supervision enriches sensor-action representations beyond action prediction. Beyond predicting which action to take, we examine whether OpenSLA captures the physiological information underlying these decisions. We evaluate this through zero-shot physiological state estimation from generated captions. As shown in Table 5, OpenSLA produces the most accurate physiological state estimates among all baselines, indicating that its generated language preserves rich information about the underlying sensor state. The gains are most pronounced in the clinical domain, where different modalities provide complementary information that needs to be effectively integrated.

What drives the strong sensor-language-action capability? The strong alignment observed above motivates us to examine what contributes to the connection between generated language and input signals. We first assess whether generated captions capture action-relevant evidence by comparing them with ground-truth action-evidence captions. As shown in Table 6, OpenSLA achieves the

Table 6: **Action explanation results.** We report BERTScore$^{\uparrow}$ (F1) between the generated captions and ground-truth action-evidence captions. OpenSLA achieves the best overall performance across all domains.

Method	Clinical		Operating Room		CGM	
	MC-MED	MIMIC-III	MOVER	VitalDB	MetaboNet	PEDAP
Multi-modal LLMs:						
GPT-5.6-LUNA	75.1	75.1	70.7	71.4	76.1	76.6
DEEPSEEK-V3.2	73.8	73.2	69.9	71.2	74.5	75.5
TS-LLM:						
OPENTSLM-SP	79.5	80.3	**81.7**	80.7	84.4	83.2
SENSORLM	77.2	78.8	67.9	69.8	76.1	76.4
Ours:						
OpenSLA-H	**82.0**	**80.9**	81.0	80.7	80.4	76.7
OpenSLA-B	81.5	78.6	81.5	**80.9**	**85.6**	**87.6**

strongest overall grounding of action decisions in sensor evidence and individual context. Notably, TS-LLMs such as OPENTSLM and SENSORLM also approach OpenSLA after adaptation on our training corpus. In contrast, zero-shot LLMs such as GPT-5.6-LUNA remain substantially behind, supporting the role of action-relevant language supervision in learning action-aligned representations.

To examine this grounding more directly, we qualitatively compare generated captions with the underlying input signals. We compare responses produced by OpenSLA, GPT-5.6-LUNA, and DEEPSEEK-V3.2 on a representative case. As shown in Fig. 4, OpenSLA generates more precise and clinically grounded descriptions than zero-shot multi-modal LLMs, suggesting that it captures richer information from the input signals.

6 WORLD KNOWLEDGE EMERGES IN SENSOR-LANGUAGE-ACTION MODEL

Semantic structure emerges in the embedding space of SLA and enables unseen action generalization. Building on the SLA alignment observed above, we want to identify whether it enables capabilities beyond conventional sensor-language modeling. We begin by examining the geometry of the learned sensor-action space. Specifically, we extract backbone representations from MIMIC-III and visualize them on PCA dimension-reduced space. As shown in Fig. 5, semantically related actions occupy nearby regions in the embedding space. For example, antibiotic medications form a compact cluster, and the unseen action Imipenem, which shares the same antibiotic semantics, is located close to it.

Motivated by this structure, we further evaluate zero-shot generalization to unseen actions. As shown in Fig. 5, OpenSLA achieves the best performance on Imipenem, which is excluded from training. Notably, it maintains a substantial margin over all baselines, including OPENTSLM after adaptation on our training data.

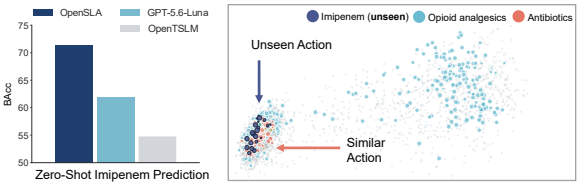

Figure 5: **Generalization to unseen concepts.**

Future-state information emerges without explicit supervision. We next examine how this semantic structure extends from discrete action categories to continuous action spaces. We extract representations and interpolate between samples with two distinct actions. For each interpolated point, we retrieve the nearest real-sample representation and use its corresponding action value. As shown in Fig. 6, the retrieved actions vary monotonically along the interpolation path, revealing a continuous organization of insulin actions in the learned representation space.

We therefore ask whether the learned representations capture not only action semantics but also the dynamics needed to predict the next sensor state, despite being trained without explicit future-state supervision. To test this, we probe whether the current representations encode information about subsequent glucose states. We construct paired samples consisting of a 2-hour CGM observation

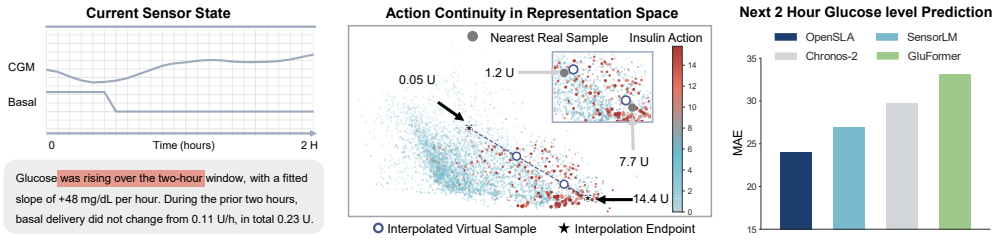

Figure 6: `OpenSLA` **exhibits emergent world-modeling capability. Left:** Current 2-hour CGM observations. **Middle:** We interpolate between representations of samples with distinct insulin actions, and retrieve the nearest real sample at each point. The corresponding actions vary monotonically along the path. **Right:** Future glucose states can be accurately read out from the learned representations.

window and summary statistics of the following 2 hours, including the mean, minimum, maximum, and temporal slope. As shown in Fig. 6, representations from `OpenSLA` enable more accurate next-state prediction under linear probing than those from competing methods. Taken together, these results suggest that world knowledge can emerge from sensor-language-action modeling.

Sensor–action knowledge transfers to an external cohort. To examine whether the learned representations generalize beyond the training cohorts, we evaluate the Clinical models on waveform-linked MIMIC-IV samples without additional fine-tuning, retaining the existing Clinical action vocabulary. As shown in Table 7, both `OpenSLA` variants outperform the adapted sensor–language baselines on necessity and category BAcc. These results suggest that `OpenSLA` learns transferable associations between sensor observations, individual context, and actions, supporting the reuse of a trained model on an external clinical cohort without cohort-specific adaptation.

Table 7: **MIMIC-IV action prediction results.** AUROC and BAcc (%). A dash denotes an unavailable continuous score.

Method	Necessity		Category	
	AUROC↑	BAcc↑	AUROC↑	BAcc↑
OpenTSLM-SP Adapt	–	69.30	–	54.33
SensorLM Adapt	–	67.70	–	50.02
`OpenSLA-B`	**81.75**	**74.40**	74.64	56.35
`OpenSLA-H`	80.52	73.00	**74.90**	**58.49**

Ablation Study. We compare `OpenSLA` with a few ablation variants. As shown in Table 15, `OpenSLA` achieves the strongest overall performance for sensor-action modeling. The largest performance gap arises from the modeling paradigm itself, with VLA-style caption-free training performing the worst, highlighting the benefit of SLA modeling with language supervision. Removing action conditioning or multi-modal fusion decoder further degrades performance, suggesting that effective integration of multi-modal context is also important for sensor-action prediction.

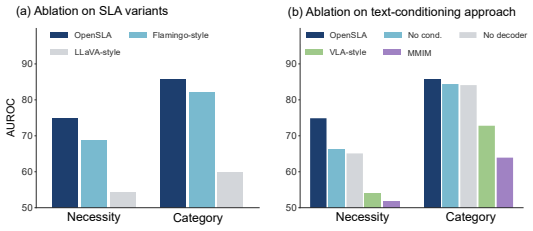

Figure 7: **Ablation study.** `OpenSLA` achieves the best among all compared variants.

7 Conclusion

We introduced **Sensor-Language-Action (SLA)** modeling, extending sensor intelligence from perceiving the world to reasoning about what to do next. SLA uses language as a semantic interface to unify sensor observations, context, and actions within a shared modeling framework. We instantiate this paradigm with a large-scale benchmark, multi-faceted language supervision, and `OpenSLA`, achieving strong performance across action prediction, state understanding, and action explanation, together with evidence grounding and zero-shot generalization to unseen actions and cohorts. Our results point to a broader role for language in sensor intelligence: not simply to describe observations, but to organize the knowledge that connects *states to actions*. We believe SLA opens a path toward general-purpose sensor models that can perceive, reason, and act through a common semantic space.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing parts of this paper, finding related work, and writing and debugging code for data preparation, model implementation, evaluation, and figure and table preparation. We reviewed and revised AI-assisted material and did not treat generated output as scientific evidence. The authors take responsibility for the final study design, analyzes, claims, results, and artifacts in this paper.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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. URL https://arxiv.org/abs/2510.15821.

DeepSeek-AI et al. DeepSeek-V3.2: Pushing the frontier of open large language models. arXiv preprint arXiv:2512.02556, 2025. URL https://arxiv.org/abs/2512.02556.

Navid Mohammadi Foumani, Chang Wei Tan, Geoffrey I Webb, and Mahsa Salehi. Improving position encoding of transformers for multivariate time series classification: Nm foumani et al. *Data mining and knowledge discovery*, 38(1):22–48, 2024.

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. URL https://proceedings.mlr.press/v235/goswami24a.html.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9192. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.723. URL https://aclanthology.org/2021.emnlp-main.723/.

Shahad Hardan, Mai A Shaaban, Jehad Abdalla, and Mohammad Yaqub. Affordable and real-time antimicrobial resistance prediction from multimodal electronic health records. *Scientific Reports*, 14(1):16464, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

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*, 3:160035, 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.

Justin Khasentino, Anastasiya Belyaeva, Xin Liu, Zhun Yang, Nicholas A Furlotte, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. A personal health large language model for sleep and fitness coaching. *Nature Medicine*, 31(10):3394–3403, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A. Xu, Winnie Chow, Martin Maritsch, Robert Jakob, Ning Wang, Juncheng Liu, Aradhana Verma, Brian Han, Daniel Seung Kim, Henry Chubb, Scott Ceresnak, Aydin Zahedivash, Alexander Tarlochan Singh 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. arXiv preprint arXiv:2510.02410, 2026. URL https://arxiv.org/abs/2510.02410.

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, 2022a. doi: 10.1038/s41597-022-01411-5.

Seung Mi Lee, Garam Lee, Tae Kyong Kim, Trang Le, Jie Hao, Young Mi Jung, Chan-Wook Park, Joong Shin Park, Jong Kwan Jun, Hyung-Chul Lee, et al. Development and validation of a prediction model for need for massive transfusion during surgery using intraoperative hemodynamic monitoring data. *JAMA network open*, 5(12):e2246637, 2022b.

Jun Li, Aaron D Aguirre, Valdery Moura Junior, Jiarui Jin, Che Liu, Lanhai Zhong, Chenxi Sun, Gari Clifford, M Brandon Westover, and Shenda Hong. An electrocardiogram foundation model built on over 10 million recordings. *Nejm ai*, 2(7):AIoa2401033, 2025a.

Sirui Li, Shuhan Xiao, Mihir Joshi, Ahmed Metwally, Daniel McDuff, Wei Wang, and Yuzhe Yang. Hearts: Benchmarking llm reasoning on health time series. *arXiv preprint arXiv:2603.06638*, 2026.

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, 2025b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Guy Lutsker, Gal Sapir, Smadar Shilo, Jordi Merino, Anastasia Godneva, Jerry R Greenfield, Dorit Samocha-Bonet, Raja Dhir, Francisco Gude, Shie Mannor, et al. A foundation model for continuous glucose monitoring data. *Nature*, 650(8103):978–986, 2026.

Girish Narayanswamy, Xin Liu, Kumar Ayush, Yuzhe Yang, Xuhai Xu, Shun Liao, Jake Garrison, Shyam Tailor, Jake Sunshine, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Samy Abdel-Ghaffar, and Daniel McDuff. Scaling wearable foundation models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yb4QE6b22f.

OpenAI. GPT-5.6 Luna. OpenAI API documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.6-luna. Accessed September 18, 2026.

Arvind Pillai, Dimitris Spathis, Fahim Kawsar, and Mohammad Malekzadeh. Papagei: Open foundation models for optical physiological signals. In *International Conference on Learning Representations*, volume 2025, pp. 48230–48261, 2025.

Muntaha Samad, Mirana Angel, Joseph Rinehart, Yuzo Kanomata, Pierre Baldi, and Maxime Cannesson. Medical informatics operating room vitals and events repository (MOVER): a public-access operating room database. *JAMIA Open*, 6(4):ooad084, 2023. doi: 10.1093/jamiaopen/ooad084.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Zitao Shuai, Zongzhe Xu, David Yang, Wei Wang, and Yuzhe Yang. OSF: On pre-training and scaling of sleep foundation models. *arXiv preprint arXiv:2603.00190*, 2026. URL https://arxiv.org/abs/2603.00190.

Harini Suresh, Nathan Hunt, Alistair Johnson, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Clinical intervention prediction and understanding with deep neural networks. In *Machine learning for healthcare conference*, pp. 322–337. PMLR, 2017.

R. Paul Wadwa, Zachariah W. Reed, Bruce A. Buckingham, Mark D. DeBoer, Laya Ekhlaspour, Gregory P. Forlenza, Melissa Schoelwer, John Lum, Craig Kollman, Roy W. Beck, and Marc D. Breton. Trial of hybrid closed-loop control in young children with type 1 diabetes. *New England Journal of Medicine*, 388(11):991–1001, 2023. doi: 10.1056/NEJMoa2210834.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better LLM agents. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 50208–50232. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24h.html.

Aaron C Weidman, Salim Malakouti, David D Salcido, Chase Zikmund, Ravi Patel, Leonard S Weiss, Michael R Pinsky, Gilles Clermont, Jonathan Elmer, Ronald K Poropatich, et al. A machine learning trauma triage model for critical care transport. *JAMA Network Open*, 8(6):e259639, 2025.

Miriam K. Wolff, Peter Calhoun, Eleonora Maria Aiello, Yao Qin, and Sam F. Royston. MetaboNet: The largest publicly available consolidated dataset for type 1 diabetes management. arXiv preprint arXiv:2601.11505, 2026. URL https://arxiv.org/abs/2601.11505.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53140–53164. PMLR, 2024. URL https://proceedings.mlr.press/v235/woo24a.html.

Justin Xu, Xi Zhang, Javid Abderezaei, Julie Bauml, Roger Boodoo, Fatemeh Haghighi, Ali Ganjizadeh, Eric Brattain, Dave Van Veen, Zaiqiao Meng, et al. Radeval: A framework for radiology text evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 546–557, 2025.

Zongzhe Xu, Aakarsh Anand, Sarah Jiang, Chuntung Zhuang, Zitao Shuai, Sriram Sankararaman, and Yuzhe Yang. Inertia-1: An open exploration of wearable motion foundation models. *arXiv preprint arXiv:2607.06617*, 2026a.

Zongzhe Xu, Zitao Shuai, Eideen Mozaffari, Ravi S. Aysola, Rajesh Kumar, and Yuzhe Yang. SleepLM: Natural-language intelligence for human sleep. arXiv preprint arXiv:2602.23605, 2026b. URL https://arxiv.org/abs/2602.23605.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, Maxwell A. Xu, Ahmed A. Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, Tim Althoff, Yun Liu, Pushmeet Kohli, Jiening Zhan, Mark Malhotra, Shwetak Patel, Cecilia Mascolo, Xin Liu, Daniel McDuff, and Yuzhe Yang. SensorLM: Learning the language of wearable sensors. arXiv preprint arXiv:2506.09108, 2025. URL https://arxiv.org/abs/2506.09108.

Table 8: **Dataset sizes used for training and action evaluation.** Training counts refer to the eligible or sampled training pool. Validation and test counts refer to the action benchmark subsets, not the full held-out candidate pools.

	Clinical		Operating Room (OR)		CGM	
Split	MC-MED	MIMIC-III	MOVER	VitalDB	MetaboNet	PEDAP
Train	90,412	16,184	113,865	48,800	255,401	16,599
Validation	403	129	238	1,000	10,000	10,000
Test	2,753	3,402	795	2,000	10,000	10,000

A Details of Datasets

A.1 Dataset Splits and Statistics

We use MC-MED (Kansal et al., 2025) and MIMIC-III (Johnson et al., 2016) for clinical care, MOVER (Samad et al., 2023) and VitalDB (Lee et al., 2022a) for operating room (OR) care, and MetaboNet (Wolff et al., 2026) and PEDAP (Wadwa et al., 2023) for continuous glucose monitoring (CGM). Both datasets in each domain contribute training samples. Table 8 reports the eligible training pools and the validation and test subsets used for action evaluation. Validation samples are used for threshold selection; these benchmark subsets are distinct from the full held-out candidate pools and caption-evaluation cohorts.

The clinical training pool contains 106,596 samples from approximately 46,000 individuals, while the operating room (OR) pool contains 162,665 samples spanning approximately 31,000 surgical cases. For CGM, we sample 272,000 training windows from MetaboNet and PEDAP, rather than training on the entire constructed corpus of approximately 2.6 million windows.

A.2 Sensor Modalities and Coverage

Waveforms and numerical measurements are complementary sensor inputs, with modality availability varying across datasets and samples. Table 9 groups the modalities by domain and lists representative measurements. In CGM, glucose traces form the waveform input, while historical basal-insulin delivery forms the numerical input.

We measure coverage as the proportion of training samples containing at least one available signal of the corresponding type, including both action-positive and action-negative samples in the denominator. This is sample-level availability, not temporal completeness or coverage of every listed modality. For example, 100% glucose coverage means that every CGM training sample contains valid glucose readings, not that every time point is observed. Valid zero-valued basal-insulin records count as available numerical measurements.

A.3 Sensor-Action Sample Construction

Input State. Each sample is anchored at a decision time T_0 and combines a historical sensor window with textual individual context. The input windows span 30 minutes for clinical care, 5 minutes for operating room (OR) care, and 2 hours for CGM. The CGM model input contains 24 five-minute slots preceding the decision slot. Text context records available individual background, presentation, and prior actions, depending on the source dataset. The recorded target action is stored separately as supervision.

Action Targets. We organize action annotations into necessity, category, and fine-grained labels where applicable. Necessity indicates whether an action of the relevant type is recorded; categories group related actions, and fine-grained labels identify specific actions within these categories. Clinical and operating room (OR) samples can have multiple category and label targets. The CGM task predicts bolus occurrence and its dose bin, without a separate fine-grained label vocabulary. Negative samples encode the absence of a target action under the dataset's annotation protocol; they do not imply that the individual's state is physiologically normal. These targets describe observed actions, rather than establishing which action would be optimal.

Table 9: **Sensor modalities and training-set coverage.** We summarize modality families with representative measurements. Coverage is the percentage of training samples containing at least one available signal of the corresponding type; it does not measure the fraction of the time window observed.

Sensor Type	Modalities	Training Coverage (%)	
Clinical		MC-MED	MIMIC-III
Waveform	ECG leads; plethysmography; respiration; arterial and other invasive-pressure waveforms.	70.8	78.8
Numerical	Vital signs (HR, RR, SpO_2, blood pressure, temperature); hemodynamics; ventilator measurements; oxygen delivery; glucose; neurological and pain scores; urine output.	99.0	42.9
Operating Room (OR)		MOVER	VitalDB
Waveform	ECG leads; plethysmography; arterial and central venous pressure; capnography; airway pressure; EEG.	25.1	100.0
Numerical	Vital signs; invasive and non-invasive blood pressure; cardiac output and related hemodynamics; ventilation and respiratory gases; BIS and EEG-derived indices; laboratory measurements; urine output.	99.6	100.0
Continuous Glucose Monitoring (CGM)		MetaboNet	PEDAP
Waveform	Glucose time series.	100.0	100.0
Numerical	Historical basal-insulin delivery.	100.0	100.0

B Multi-Faceted Caption Construction

We construct caption targets through structured fact extraction followed by rule-based selection and text rendering. For each decision time T_0, the pipeline extracts findings from the preceding sensor window and available individual context. Intermediate records retain the measurement identity, numerical summary, temporal information, and evidence source. These records are organized into four complementary caption facets. Recorded actions guide evidence selection where applicable, but do not establish the presence of a physiological finding.

B.1 Observational Caption Construction

Observational captions describe individual measurements and waveform channels. For numerical observations, we summarize available values using the median, observed range, and observation count, and associate these summaries with measurement-specific state descriptors. Sparse measurements retain their observed values and sample counts without implying a continuous trajectory. For waveforms, channel-specific extractors produce rate, periodicity, amplitude, and pattern descriptors where supported. Derived quantities are identified as waveform proxies rather than treated as interchangeable with charted measurements.

The selected findings are rendered with their measurement names and numerical summaries. Quality-limited waveforms and incompatible proxy estimates receive qualified availability descriptions instead of unsupported rate or morphology statements. For CGM, the glucose summary includes its median and range, fitted slope, and latest available reading; additional range-occupancy and variability descriptions are included when their corresponding conditions are met.

B.2 Relational Caption Construction

Relational captions describe temporal structure and relationships between physiological measurements. Temporal descriptions are derived from the extracted trajectories, distinguishing directional changes from persistent or relatively unchanged states. Trend statements require sufficient observations. Where historical event records are available, the captions also summarize their timing and duration, such as earlier carbohydrate entries, insulin delivery, or basal-rate changes in CGM.

Cross-signal descriptions use predefined, physiologically related channel pairs rather than unrestricted pairwise correlations. Examples include numerical heart rate with ECG rate proxies and

respiratory rate with respiration-waveform patterns. A joint statement is generated only when the relevant observations are available and satisfy the corresponding compatibility and quality rules. This preserves the distinction between observing two channels and establishing support for a shared physiological finding.

B.3 Global Caption Construction

Global captions combine available individual context with selected physiological findings. Context includes recorded background information, presentation, and relevant history. The state-summary component selects among predefined state descriptions using their supporting facts, rather than freely inferring a diagnosis from the complete record.

For Clinical and operating room data, candidate states are ranked by the support provided by available findings, and the selected description is accompanied by its principal supporting evidence. CGM uses predefined combinations of glucose dynamics and historical insulin or carbohydrate information to identify state descriptions. If no eligible state is supported, the caption reports limited evidence or the absence of a selected state, rather than assuming normal physiology.

B.4 Action Caption Construction

Action captions connect sensor observations and individual context with action-relevant evidence. We use predefined matching rules to select relevant findings from the extracted facts and summarize how they support or qualify an action-related interpretation. Where applicable, recorded action annotations guide this selection, while the underlying evidence is drawn from pre-decision observations. The resulting descriptions express evidence associations rather than causal explanations or treatment recommendations.

Finally, the selected facts are verbalized using sentence templates and phrase banks that vary wording while retaining measurement identities, values, and evidence qualifications. Complete constructed captions are provided in Appendix F.1, and the corresponding template pools are listed in Appendix F.2.

C Method Implementation Details

C.1 OpenSLA Implementation

Sensor and Language Encoding. For waveform encoding, we use a shared single-channel DINO encoder pretrained using the OSF codebase (Shuai et al., 2026). The encoder remains frozen during SLA training. For CGM, a trainable one-dimensional convolutional patchifier produces the corresponding temporal tokens. The numerical encoder is a trainable Transformer encoder. Each observation is converted into a token by combining an MLP-based value embedding with measurement-identity, source, and temporal embeddings. Per-measurement summary tokens are constructed through a separate MLP and appended to the observation tokens. Self-attention contextualizes these tokens across measurements and time, producing features for individual observations and measurement-level summaries.

Structured Action Heads. All three heads use the backbone hidden state at the final input-context position, denoted by $\mathbf{z}_i \in \mathbb{R}^d$ for sample i, where d is the language hidden width. This position precedes the target caption and therefore cannot attend to its tokens. *Necessity head.* A linear layer predicts independent binary necessity scores, supervised with binary cross-entropy $\mathcal{L}_{\text{nec}}$. *Category head.* A second linear layer predicts conditional category probabilities. Each probability is multiplied by its associated necessity probability to obtain the final category score. Categories are supervised with positive-class-weighted binary cross-entropy $\mathcal{L}_{\text{cat}}$, allowing multiple categories to be selected simultaneously. *Fine-grained label head.* Candidate descriptors are encoded by averaging their token embeddings from the language backbone. Separate two-layer MLPs f_q and f_d, with GELU activations, project the decision state and descriptor embeddings into a shared ranking space. Their temperature-scaled cosine similarity defines the candidate score:

$$s_{i\ell} = \frac{\cos(f_q(\mathbf{z}_i), f_d(\bar{\mathbf{e}}_\ell))}{\tau}, \tag{1}$$

where $\bar{\mathbf{e}}_\ell$ is the pooled descriptor embedding for label ℓ and $\tau > 0$ is the temperature. The ranking loss $\mathcal{L}_{\text{label}}$ is softmax cross-entropy over each positive label and sampled negatives from the same category, excluding other positive labels of that sample. It is averaged over positive labels with valid negatives. Together, the three head losses expand $\mathcal{L}_{\text{act}}$ in the training objective of Section 4 as

$$\mathcal{L}_{\text{act}} = \lambda_{\text{nec}}\mathcal{L}_{\text{nec}} + \lambda_{\text{cat}}\mathcal{L}_{\text{cat}} + \lambda_{\text{label}}\mathcal{L}_{\text{label}}. \tag{2}$$

Hierarchical Channel State Memory. The HCSM pathway uses two query stages. Local queries aggregate encoder features within each channel–time group, while temporal queries aggregate the resulting local memories across time within the same channel. Within each branch, local query parameters are shared across channel–time groups, and temporal query parameters are shared across channels. Channel embeddings condition both stages, while time embeddings additionally condition local queries. Invalid input positions are masked during attention, and memories from empty groups are excluded. Both local and temporal outputs are retained and concatenated with the global memory. Each query stage comprises multi-head cross-attention followed by a residual feed-forward block. Waveform and numerical branches instantiate this structure with independent parameters, feature widths, and query budgets. For numerical inputs, measurement identities define the channels, and observations are assigned to time groups using their relative timestamps. Per-measurement summary tokens bypass the query hierarchy and are appended separately to the numerical memory.

Multi-modal Fusion Decoder. Action tokens $\mathbf{P}$ encode predicted necessity and selected categories using learned necessity embeddings, category descriptors, and projected confidence scores. The predicted probabilities are detached during training. In the hierarchical configuration, separate waveform and numerical selectors use $\mathbf{z}_i$ and pooled $\mathbf{P}$ to retrieve pre-compression features. Waveform candidates are projected patch features with channel/time embeddings; numerical candidates are observation tokens, excluding summaries. A bounded number of candidates per channel or measurement are selected, projected into the language hidden space, and weighted by learned retrieval scores to form $\mathbf{S}$.

The action branch f_a and sensor-evidence branch f_s each contain cross-attention, a feed-forward block, and an output projection. Both query the original caption states $\mathbf{h}$, independently reading action tokens $\mathbf{P}$ and selected sensor tokens $\mathbf{S}$, respectively; the base configuration uses only the action branch. Refined caption states $\mathbf{h}'$ feed the backbone's original language output head. Training uses ground-truth caption prefixes and predicted actions, with token-weighted cross-entropy:

$$\mathcal{L}_{\text{text}} = -\frac{\sum_{i,t} w_{it} \log p_{it}(y_{it})}{\sum_{i,t} w_{it}}. \tag{3}$$

Here $p_{it}(y_{it})$ is the probability of target token y_{it} for sample i at position t, and w_{it} is its loss weight, set to zero outside caption targets. Inference conditions on previously generated tokens.

C.2 Baseline and Ablation Details

Supervised Baselines. We report ConvTrans and MMIM (Han et al., 2021) in the main action tables. *MMIM.* We adapt MMIM to combine sensor recordings with language context using a trainable CNN and frozen BERT, respectively. Their representations are fused through an MLP and passed to task-specific action classifiers. Training retains the mutual-information and contrastive objectives alongside supervised action prediction, without caption generation.

Sensor-only baselines. ConvTrans is a CNN–Transformer trained from scratch (Foumani et al., 2024), combining convolutional feature extraction with temporal self-attention. We also include frozen Chronos-2 (Ansari et al., 2025) and Moirai (Woo et al., 2024) encoders with supervised action predictors, and an MLP for CGM. For the foundation-model baselines, sensor sequences are encoded into pretrained representations and pooled before action prediction; the pretrained encoders remain frozen. These sensor-only baselines do not receive language context.

Zero-Shot and API Baselines. *OpenTSLM.* We evaluate published Llama-3.2-1B OpenTSLM (Langer et al., 2026) checkpoints without task-specific fine-tuning, retaining their original sensor encoders and language backbones. Clinical and operating room (OR) evaluations use ECG-pretrained checkpoints and an ECG-prioritized waveform channel, falling back to another available waveform when ECG is absent. The latest contiguous segment is retained within the

native input budget; numerical time series are not passed to the encoder. For CGM, M4-pretrained checkpoints receive the complete historical glucose and basal-insulin sequences. Action predictions are generated directly in a constrained JSON format using the provided action vocabulary, rather than obtained through candidate likelihood scoring.

API-based LLMs. We evaluate GPT-5.6-Luna (OpenAI, 2026) and DeepSeek-V3.2 (DeepSeek-AI et al., 2025) through a CodeAct-style interface (Wang et al., 2024). Instead of serializing long sensor recordings into the prompt, the models execute Python code to inspect pre-decision measurements and compute relevant summaries. Language context and tool outputs provide the observations for subsequent action prediction and evidence-caption generation. Action prompts specify the available categories and, for fine-grained evaluation, the label vocabulary of the given reference category. We evaluate the returned discrete selections without converting them into probability scores.

Adapted Sensor–Language Baselines. *OpenTSLM.* To accommodate the full sensor histories used by `OpenSLA`, we replace OpenTSLM's native sensor encoder with our domain-specific sensor encoders. We retain the original SP and Flamingo fusion architectures and initialize their remaining components from the published Llama-3.2-1B checkpoints (Langer et al., 2026), using ECG-pretrained weights for clinical and operating room (OR) tasks and M4-pretrained weights for CGM. *SensorLM.* We implement a SensorLM-style CoCa architecture (Zhang et al., 2025) using the same sensor encoders, a unimodal text encoder, and a multi-modal decoder that cross-attends to sensor features. The pretrained DINO encoder remains frozen where applicable. Training combines sensor–text contrastive learning with autoregressive generation.

Ablation Variants. We construct ablations from the `OpenSLA-B` configuration with Qwen-3.5-2B. *Caption supervision and conditioning.* The action-only variant removes caption generation and its training loss while retaining the action predictors. The no-decoder variant retains caption supervision but generates text directly from the language backbone without the residual fusion decoder. The no-action-conditioning variant retains the decoder but replaces its predicted-action inputs with a learned action-independent token. These variants distinguish the contributions of caption supervision, the additional decoder, and explicit action conditioning.

Language-backbone adaptation. We additionally compare two alternatives to LoRA-based adaptation. The LLaVA-style variant (Liu et al., 2023) freezes the language backbone and introduces sensor features through trainable projectors as input tokens. The Flamingo-style variant (Alayrac et al., 2022) instead integrates sensor features through trainable gated cross-attention modules within the frozen backbone. Both retain the base model's sensor encoders, action heads, and caption supervision, with sensor-side and task-specific modules remaining trainable except for the pretrained DINO encoder.

C.3 Training Details

Table 10 summarizes optimization settings and effective batch sizes including gradient accumulation. The SensorLM-style text encoder and decoder are randomly initialized rather than loaded from pretrained SensorLM weights.

D Task Setup

D.1 Action Prediction

We evaluate action prediction at three levels: necessity, category, and fine-grained label. In Clinical and operating room (OR) tasks, necessity indicates whether any target action occurs, while categories and fine-grained labels specify the action. CGM necessity indicates a bolus intervention, and its categories correspond to dose bins; there is no separate fine-grained label task. Text-generating baselines are evaluated by mapping their generated actions to the same target vocabulary.

We report AUROC from continuous prediction scores and balanced accuracy (BAcc), defined as the mean of sensitivity and specificity. For the main action tables, thresholds maximize validation BAcc separately for each binary target, retaining the calibration scope of each evaluated model. Fine-grained labels are evaluated within ground-truth-positive category cohorts, averaging eligible

Table 10: **Training settings for OpenSLA and adapted sensor–language baselines.** Batch sizes include gradient accumulation. LoRA settings apply to OpenSLA and OpenTSLM-SP, not OpenTSLM-Flamingo. A dash denotes a setting that does not apply.

Setting	OpenSLA	OpenTSLM	SensorLM
Optimization:			
Optimizer	AdamW	AdamW	AdamW
Peak learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}
Weight decay	0.01	0.01	0.01
Warmup fraction	0.10	0.10	0.10
Learning-rate decay	Linear	Linear	Linear
Gradient clipping norm	0.5	0.5	0.5
LoRA rank / α / dropout	16 / 32 / 0.05	16 / 32 / 0.05	–
Effective Batch Size:			
Clinical	16	16	64
Operating room (OR)	16	16	64
CGM	16	16	256

labels within each category and then across categories. This conditional label metric is distinct from end-to-end action accuracy.

D.2 Caption-Based Evaluation

Sensor-state understanding. We extract physiological quantities from generated captions and compare them with reference statistics computed from the observed history. Table 11 specifies the quantity, statistic, and readout for each task. We compute mean absolute error (MAE) on reference-eligible samples with a valid extracted prediction, using the reference unit.

Action-evidence agreement. We compare generated text with the action-evidence portion of the reference caption using BERTScore F1. We use `distilbert-base-uncased`. The reported result is the mean per-sample F1, multiplied by 100.

D.3 Unseen Concept Classification

To examine generalization to unseen fine-grained actions within a known category, we construct a balanced binary classification set from MIMIC-III. The target is Imipenem/Cilastatin, a single antibiotic label absent from the supervised action training targets, while the antibiotic category is seen during training. The set contains:

- **Positive samples:** Individuals with Imipenem/Cilastatin administration in the five-minute action-target window.
- **Negative samples:** Individuals with neither an annotated action nor a recorded MetaVision medication or procedure event in the same target window.

Zero-Shot Action Scoring. `OpenSLA` represents candidate actions through textual descriptors rather than label-specific classifier weights. At inference, we encode the unseen descriptor, `antibiotic: Imipenem/Cilastatin`, by averaging its language-backbone token embeddings and applying the trained descriptor projection. Its cosine similarity with the projected multi-modal decision representation provides the fine-grained action score. All model parameters remain fixed; no additional training on the unseen label is required.

D.4 External-Cohort Evaluation

We evaluate the Clinical models on waveform-linked MIMIC-IV samples without further adaptation to this cohort. Each sample provides pre-decision sensor history and language context, with actions mapped to the Clinical target vocabulary. We report necessity and category AUROC and BAcc; text-only predictions contribute BAcc but not AUROC. This experiment measures transfer to another MIMIC release and cohort.

Table 11: **Caption evaluation tasks and scoring rules.** Definitions for the sensor-state and action-evidence tasks reported in the main results.

Quantity / task	Reference statistic	Prediction readout	Metric
Clinical: 30-minute sensor history			
Cardiac index	Window median ($L/min/m^2$)	Extract an explicit cardiac-index median.	MAE↓
ECG-II rate	ECG-derived rate proxy (beats/min)	Extract the rate attributed to ECG lead II.	MAE↓
RESP rate	Respiration-waveform rate proxy (breaths/min)	Extract a rate attributed to the RESP waveform.	MAE↓
Operating room: 5-minute sensor history;			
Respiratory rate	Window median (breaths/min)	Extract an explicit median of measured RR.	MAE↓
Temperature	Window median (°C)	Extract an explicit temperature median.	MAE↓
End-tidal CO_2	Window median (mmHg)	Extract an explicit $ETCO_2$ median.	MAE↓
CGM: 2-hour pre-decision history			
Basal rate	Mean recorded basal-insulin rate (U/h)	Extract a mean/average basal rate.	MAE↓
Carbohydrate amount	Amount of the latest recorded carbohydrate event (g)	Extract an amount paired with a pre-decision recency in the preceding 120 minutes, and choose the latest unambiguous event.	MAE↓
Action-evidence text: all six development datasets			
Action-evidence agreement	Ground-truth action-evidence text	Compare generated captions, against the same reference evidence.	BERTScore F1↑

E Additional Results

We provide additional results on sensor–language adaptation, caption supervision, generalization, and the information retained in learned representations. The operating room (OR) ablations use `OpenSLA-B` as the reference model and retain the existing test cohorts and checkpoints; training steps differ across variants.

E.1 Sensor–Language Adaptation

Adapting the backbone versus external fusion. We compare direct LoRA adaptation with two alternatives: LLaVA-style training, which freezes the language backbone and learns sensor projectors, and Flamingo-style training, which introduces trainable cross-attention into a frozen backbone. These variants retain the base model's action heads and caption supervision. Table 12 shows that direct adaptation has the highest necessity and category AUROC on both MOVER and VitalDB, while Flamingo-style fusion consistently improves over projector-only training.

Table 12: **Sensor–language adaptation.** Necessity and category AUROC (%) on operating room datasets.

Method	MOVER		VitalDB	
	Necessity	Category	Necessity	Category
LLaVA-style	51.5	66.4	54.4	60.0
Flamingo-style	82.9	76.1	68.9	82.1
`OpenSLA-B`	**83.4**	**80.4**	**74.9**	**85.9**

Thresholded decisions. Table 13 complements AUROC with balanced accuracy (BAcc). Thresholds maximize per-class BAcc on the pooled OR validation set, using 0.5 when a class lacks positive or negative validation examples. Category results are macro-averaged; fine-label results average labels within ground-truth-positive categories and then average categories. `OpenSLA-B` leads category BAcc on both datasets and fine-label BAcc on VitalDB, tying Flamingo-style fusion on MOVER fine labels. Flamingo-style fusion has higher necessity BAcc, showing that the AUROC advantage does not imply uniformly better decisions at the selected thresholds.

Table 13: **Sensor–language adaptation with BAcc.** Balanced accuracy (%); Nec., Cat., and Label denote necessity, category, and fine-grained label.

Method	MOVER			VitalDB		
	Nec.	Cat.	Label	Nec.	Cat.	Label
LLaVA-style	53.4	59.8	54.9	52.3	56.7	51.1
Flamingo-style	**79.4**	68.3	**66.4**	**62.9**	77.8	74.6
`OpenSLA-B`	74.2	**69.0**	**66.4**	59.0	**79.5**	**78.9**

E.2 Caption Supervision and Conditioning

Which caption facets support action prediction? We compare observational and relational supervision alone with all four caption facets, adding global context and action-relevant evidence together. As shown in Table 14, the full-caption setting improves necessity and category AUROC on both datasets. The largest difference is in VitalDB necessity, increasing from 65.9 to 74.9. This comparison supports the combined contribution of context and action-oriented descriptions beyond signal summaries; it does not separate the individual effects of global and action captions.

Table 14: **Caption-supervision ablation.** Necessity and category AUROC (%). Obs.: observational; Rel.: relational.

Caption Facets				MOVER		VitalDB	
Obs.	Rel.	Global	Action	Necessity	Category	Necessity	Category
✓	✓	✗	✗	79.2	76.4	65.9	82.6
✓	✓	✓	✓	**83.4**	**80.4**	**74.9**	**85.9**

Is caption supervision alone sufficient? We next distinguish the presence of language supervision from the mechanism used to connect it to action prediction. The action-only variant removes caption generation and its loss. The no-action-conditioning variant retains the decoder but removes predicted-action conditioning; the no-decoder variant retains caption supervision and generates text directly from the backbone. Table 15 shows that the full model has the highest AUROC across all four endpoints. Action-only training is lowest throughout, while both caption-generating ablations improve over it. Explicit action conditioning and the fusion decoder provide further gains in these comparisons.

Table 15: **Caption generation and conditioning.** Necessity and category AUROC (%) for variants of `OpenSLA-B`.

Method	MOVER		VitalDB	
	Necessity	Category	Necessity	Category
Action only	78.1	72.1	54.1	72.8
No cond.	80.4	77.2	66.3	84.4
No decoder	78.3	76.4	65.1	84.1
`OpenSLA-B`	**83.4**	**80.4**	**74.9**	**85.9**

E.3 Generalization

An unseen action within a known category. We evaluate Imipenem/Cilastatin, a single fine-grained antibiotic label absent from the supervised action targets. `OpenSLA` scores the new label through its textual descriptor without further model training. The decisions reported in Table 16 also apply the necessity and antibiotic-category gates. Text-generating baselines are evaluated by target presence in their final actions. `OpenSLA-H` achieves the highest F1 and BAcc.

Table 16: **Unseen-action detection on MIMIC-III.** Imipenem/Cilastatin versus no action; F1 and BAcc (%).

Method	Imipenem	
	F1$^{\uparrow}$	BAcc$^{\uparrow}$
OpenTSLM-SP Adapt	17.4	54.8
GPT-5.6-Luna	42.9	61.9
DeepSeek-V3.2	34.5	54.8
`OpenSLA-B`	52.6	57.1
`OpenSLA-H`	**70.0**	**71.4**

E.4 Future-State Readout

We test whether current-state representations retain information about subsequent glucose dynamics. The MetaboNet study pairs a two-hour observation window with glucose mean, minimum, maximum, and slope over the following two hours, using 9,876 pairs from 128 individuals. The backbone representations remain frozen while ridge readouts are fitted with individual-grouped nested cross-validation: five outer folds and three inner folds for regularization selection. No handcrafted numerical summaries are concatenated with the embeddings. The Chronos-2 row instead uses its native zero-shot forecasts; the GluFormer row uses the CRUISE reproduction.

Table 17 reports glucose-summary MAE in mg/dL and slope MAE in mg/dL/min. The `OpenSLA` variants have lower error on mean, minimum, and maximum glucose than the listed baselines. Slope errors are close: `OpenSLA`, OpenTSLM-SP, and SensorLM all round to 0.39. These results concern a trained readout of frozen representations, not zero-shot generation of future glucose values.

Table 17: **Future-state readout on MetaboNet.** MAE for the next two hours: mean, minimum, maximum, and slope.

Method	Mean$^{\downarrow}$	Min$^{\downarrow}$	Max$^{\downarrow}$	Slope$^{\downarrow}$
OpenTSLM-SP	26.5	24.2	30.6	0.39
SensorLM	26.9	24.6	30.5	0.39
Chronos-2 (zero-shot)	29.7	29.3	32.8	0.48
GluFormer	33.2	30.0	38.7	0.41
`OpenSLA-B`	24.0	**22.3**	28.0	0.39
`OpenSLA-H`	**24.0**	22.5	**27.7**	**0.39**

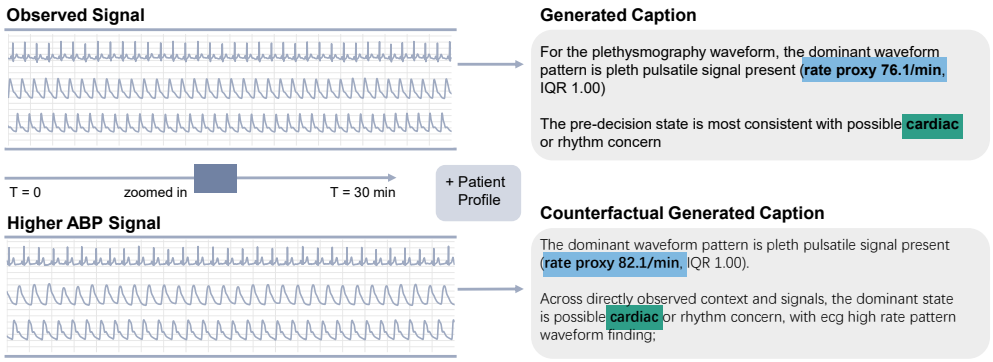

Figure 8: **Caption responses to sensor substitution.** Patient context is held fixed while the supplied sensor recording changes.

E.5 Controlled Sensor Substitution

We hold individual context fixed and replace the sensor recording with a higher-ABP example. In Fig. 8, the generated Pleth rate proxy changes from 76.1 to 82.1/min, and the state description adds a high-rate ECG finding. The retained context therefore does not fully determine the caption. This qualitative example illustrates sensitivity to sensor substitution, rather than validating a physiological intervention effect or establishing the accuracy of every generated quantity.

F Additional Examples

F.1 Multi-faceted Caption Examples

Examples Captions Generated by Our Data Pipeline

```
================================== Example 1 ==================================

OBSERVATIONAL: Numeric:  1-minute heart-rate variability was evidence of hrv
abnormal (1min_HRV median 10, range 5-32, n=19) and persistently outside the
expected range.  The heart rate series indicates tachycardia and is persistently
outside the expected range (HR median 110, range 107-120, n=19).  Waveform:  The
Pleth trace did not meet the quality threshold for morphology or rate-proxy
statements.  Because Resp produced incompatible rhythm or periodicity proxies, it
was excluded from the waveform conclusion.  Lead II waveform evidence includes
tachycardic ECG waveform proxy (rate proxy 108.8/min, IQR 0.20).

RELATIONAL: Temporal:  Temporal pattern:  short-window HRV is abnormal with
little recent change across the current decision window; evidence of hrv
abnormal.  Near the decision point, heart-rate channel shows elevated heart rate
and is abnormal but not changing much.  Cross-signal:  elevated cardiac rate
across numeric and ECG evidence is present across the corresponding signal
channels.

GLOBAL: Context:  Before the decision point, the contextual record contains
reported shortness of breath at presentation.  State summary:  Across directly
observed context and signals, the dominant state is possible cardiac or rhythm
concern finding, with hrv abnormal context; rapid cardiac rate; ecg high rate
pattern waveform finding.
```

```
ACTION: elevated heart-rate pattern (HR median 110, range 107-120, n=19); trend:
abnormal but not changing much; elevated cardiac rate across numeric and ECG
evidence; cardiac waveform channel showed high-rate Lead II periodicity (rate
proxy 108.8/min, IQR 0.20) is relevant to circulatory support, although the
available evidence is indirect or incomplete.

================================== Example 2  ==================================


OBSERVATIONAL: Numeric:  Across the current observation interval, perfusion
signal shows poor perfusion signal; Perf median 1, range 1-1, n=30; persistently
outside the expected range.  For respiratory rate, the dominant finding is
elevated respiratory rate; RR median 24, range 17-29, n=30.  Waveform:  The Pleth
trace did not meet the quality threshold for morphology or rate-proxy statements.
The rate or periodicity proxies from Resp were internally inconsistent, so that
channel was not used for waveform interpretation.  ECG channel summary:  ecg rate
proxy normal; rate proxy 81.2/min, IQR 0.10.

RELATIONAL: Temporal:  Temporal pattern:  Perf is sustained without a clear
directional shift across the half-hour signal window; poor perfusion signal.
Over the monitored interval, RR is abnormal but not changing much with elevated
respiratory rate.

GLOBAL: Context:  Available contextual evidence includes reported neurologic at
presentation.  State summary:  The pre-decision state is most consistent with
possible hemodynamic instability signal; key evidence is reduced peripheral
perfusion proxy; weak Pleth waveform.

ACTION: A possible reason for the infectious diagnostic testing workup is the
pre-T0 finding tachypnea (RR median 24, range 17-29, n=30); trend:  persistently
outside the expected range; perfusion concern (Perf median 1, range 1-1, n=30);
trend:  persistently outside the expected range; peripheral pulse waveform showed
weak pulse waveform amplitude (IQR 0.00); no direct note statement confirms it.

================================== Example 3  ==================================


OBSERVATIONAL: Numeric:  The heart rate series indicates tachycardia and is
abnormal with little recent change (HR median 107, range 102-111, n=17).
Waveform:  Because Pleth was only partly usable in this window, no morphology or
rate inference was drawn from it.  The available Resp trace was present, but its
proxy features disagreed, so no single morphology or rate interpretation was
reported.  ECG Lead II: nonuniform cardiac waveform proxy; rate proxy 99.9/min,
IQR 0.35.

RELATIONAL: Temporal:  Over the monitored interval, pulse rate is outside the
expected range without a strong trend with rapid cardiac rate.

GLOBAL: Context:  Presentation context includes chief complaint of hypertension
or hypotension.  State summary:  Across directly observed context and signals,
the dominant state is possible cardiac or rhythm concern context, with elevated
heart rate; ECG rhythm irregularity proxy.

ACTION: Observed signals and context include increased heart rate (HR median 107,
range 102-111, n=17); trend:  persistently outside the expected range;
```

```
plethysmography channel showed low-amplitude Pleth proxy (IQR 0.00); respiratory
signal showed resp irregular breathing proxy (rate proxy 14.3/min, IQR 1.29);
they provide broad clinical background for blood-pressure or volume-status
management without asserting a direct treatment indication.

================================== Example 4  ==================================

OBSERVATIONAL: Numeric:  For heart-rate channel, the dominant finding is elevated
heart-rate pattern; HR median 100, range 95-108, n=30.  Waveform:  Proxy
interpretation was deferred for Pleth because the available waveform was
quality-constrained.  Waveform availability for Resp was adequate, yet the
derived proxies were discordant and therefore not interpreted as a single
finding.  The cardiac waveform channel shows variable ECG waveform periodicity;
rate proxy 95.2/min, IQR 0.40.

RELATIONAL: Temporal:  Temporal pattern:  heart rate is abnormal but not changing
much across the current observation interval; elevated heart-rate pattern.

GLOBAL: Context:  Pre-decision context includes arrival complaint includes psych
or substance; reported medication or refill at presentation.  State summary:  The
pre-decision state is most consistent with possible cardiac or rhythm concern
signal; key evidence is elevated heart-rate pattern; irregular ECG periodicity
proxy.

ACTION: A possible reason for the evaluation of altered mental status or
neurologic concern workup is the pre-T0 finding heart rate above expected range
(HR median 100, range 95-108, n=30); trend:  outside the expected range without a
strong trend; pulse-ox waveform showed weak pulse waveform amplitude (IQR 0.00);
Resp channel showed resp irregular breathing proxy (rate proxy 17.0/min, IQR
0.99); no direct note statement confirms it.

================================== Example 5  ==================================

OBSERVATIONAL: Numeric:  The heart rate has limited charting that supports
elevated heart-rate pattern (HR median 101, range 101-101, n=1); this is based on
one pre-decision HR reading, without a reliable trajectory.  rapid breathing
pattern appears in one pre-decision RR reading (RR median 23, range 23-23, n=1);
no temporal trend is assigned.  Waveform:  ECG channel summary:  ECG high-rate
proxy; rate proxy 100.0/min, IQR 1.00.  For Plethysmography waveform, the
dominant waveform pattern is uneven Pleth pulsatility (rate proxy 91.3/min, IQR
1.00).

GLOBAL: Context:  Before the decision point, the contextual record contains
presentation mentions bleeding or gi bleed.  State summary:  Across directly
observed context and signals, the dominant state is possible infection or sepsis
concern signal, with rapid cardiac rate; RR above expected range.

ACTION: A secondary signal for symptom-control intervention is elevated
heart-rate pattern (HR median 101, range 101-101, n=1; one pre-decision HR
reading); no stronger domain-specific fact was selected.
```

```
================================== Example 6  ==================================

OBSERVATIONAL: Numeric:  Basal delivery was zero throughout the prior two hours.
Waveform:  Glucose during the preceding two hours stayed within 71-100 mg/dL,
median 88 mg/dL, inside the 70-180 mg/dL range.  In the two-hour window, the
fitted slope of sensor glucose was +6 mg/dL per hour; glucose was flat, within 10
mg/dL per hour.  The last reading before the decision was 94 mg/dL, 5 minutes
before it.

RELATIONAL: Temporal:  Over the two-hour window before the decision, no bolus,
carbohydrate entry, small delivery run or basal change was logged.

GLOBAL: Context:  Patient context:  48 yr old, a male patient, 104 kg body
weight, height of 170 cm, body mass index 36.0 kg/m2, duration of diabetes 38
years, automated insulin delivery.  State summary:  Pre-decision state:  readings
in 70-180 mg/dL throughout, changing under 10 mg/dL per hour.

ACTION: Findings on the side of no bolus:  pump basal of zero entering the
decision.

================================== Example 7  ==================================

OBSERVATIONAL: Numeric:  No basal delivery was recorded in the prior two hours.
Waveform:  Glucose over the two-hour observation window had a median of 97 mg/dL,
in the target range of 70-180 mg/dL, and went from a low of 78 to a high of 113
mg/dL. Glucose over the two-hour window before the decision:  decreasing, -18
mg/dL per hour.  Glucose at the last reading, 5 minutes before the decision, was
78 mg/dL.

RELATIONAL: Temporal:  In the two-hour lookback:  no bolus, carbohydrate entry,
small delivery run or basal change.

GLOBAL: Context:  Patient profile:  28 yr old, a male patient, a weight of 58 kg,
175 cm tall, BMI equal to 18.8 kg/m2, 11 years since diagnosis, automated insulin
delivery (AID). State summary:  Summary of the state before the decision:  none
of the defined states applied; evidence was limited.

ACTION: Against a bolus:  no pump basal in the slot before the decision.

================================== Example 8  ==================================

OBSERVATIONAL: Numeric:  Basal insulin stayed at zero for all of the last two
hours.  Insulin on board, slot before the decision:  0.13 U. Waveform:  Sensor
glucose in the preceding two hours:  median 114 mg/dL, in the 70-180 mg/dL target
range, lowest 101 and highest 122 mg/dL. Glucose over the two hours before the
decision:  stable (under 10 mg/dL per hour either way), -8 mg/dL per hour.  The
last CGM value was 101 mg/dL, 5 minutes before the decision.

RELATIONAL: Temporal:  No timed events (bolus, carbohydrate, small delivery run,
basal change) in the last two hours.
```

```
GLOBAL: Context:  Known patient attributes:  aged 43, male, a closed-loop system.
State summary:  State before the decision:  a 70-180 mg/dL trace with a slope
below 10 mg/dL per hour.

ACTION: Findings opposing an insulin bolus:  zero basal on the pump just before
the decision.

================================== Example 9  ==================================


OBSERVATIONAL: Numeric:  Basal insulin stayed at zero for all of the two-hour
window.  Waveform:  CGM readings in the last two hours spanned 88-136 mg/dL, with
a median of 117 mg/dL, neither below 70 nor above 180 mg/dL. Over the two-hour
period before the decision, glucose was dropping, at -20 mg/dL per hour.  The
reading closest to the decision was 88 mg/dL, 5 minutes earlier.

RELATIONAL: Temporal:  The record lists no bolus, carbohydrate entry, small
delivery run or basal change for the two-hour period before the decision.

GLOBAL: Context:  Patient record:  35 years old, of female sex, body weight 54
kg, 157 cm tall, BMI at 21.9 kg/m2, living with diabetes 30 years, automated
insulin delivery.  State summary:  Overall state before the decision:  no
selected glycaemic state; direct evidence was limited.

ACTION: Pre-decision evidence against an insulin bolus:  pump basal at zero just
before the decision.

================================== Example 10 ==================================


OBSERVATIONAL: Numeric:  A constant basal rate of 0.90 U/h over the two hours up
to the decision delivered 1.80 U. Insulin on board immediately before the
decision was 0.01 U. Waveform:  Median sensor glucose over the two-hour window
was 150 mg/dL, in the target range of 70-180 mg/dL; readings 142-153 mg/dL. For
the two-hour period before the decision, glucose had a fitted slope of -5 mg/dL
per hour and was flat (less than 10 mg/dL per hour).  Latest glucose:  142 mg/dL
at 5 minutes before the decision.

RELATIONAL: Temporal:  No timed events (bolus, carbohydrate, small delivery run,
basal change) in the two hours prior to the decision.

GLOBAL: Context:  Patient details on record:  aged 66, recorded as male, weighing
108 kg, a height of 185 cm, computed BMI 31.4 kg/m2, 26 years of diabetes, an
insulin pump with CGM. State summary:  Overall glycaemic state:  glucose kept
inside 70-180 mg/dL, slope under 10 mg/dL per hour.

ACTION: No finding for or against an insulin bolus was selected from the observed
window.
```

F.2 TEMPLATE POOL

OBSERVATIONAL

```
NUMERIC_CAPTION_TEMPLATES = [
  "{modality_phrase} was {numeric_finding} ({stat_phrase}) and {trend_phrase}.",
  "Across {window_phrase}, {modality_phrase} shows {numeric_finding}; {stat_phrase};
{trend_phrase}.",
  "The {modality_phrase} series indicates {numeric_finding} and is {trend_phrase}
({stat_phrase}).",
  "Recorded {modality_phrase} values support {numeric_finding}; {stat_phrase};
{trend_phrase}.",
  "For {modality_phrase}, the dominant finding is {numeric_finding}; {stat_phrase}.",
  "{numeric_finding} is present in the {modality_phrase} series and is {trend_phrase}.",
]
NUMERIC_OBSERVATION_ONLY_TEMPLATES = [
  "Observed {modality_phrase} during {window_phrase} ({stat_phrase}).",
  "The current {modality_phrase} summary is {stat_phrase}.",
  "Recorded {modality_phrase} values in {window_phrase} are summarized as {stat_phrase}.",
]
NUMERIC_SPARSE_TEMPLATES = [
  "{modality_phrase} has limited charting that supports {numeric_finding} ({stat_phrase});
this is based on {sparse_observation_phrase}, without a reliable trajectory.",
  "{numeric_finding} appears in {sparse_observation_phrase} ({stat_phrase}); no temporal
trend is assigned.",
  "Evidence from {sparse_observation_phrase} is consistent with {numeric_finding}
({stat_phrase}); the values are retained but temporal detail is withheld because
observations are sparse.",
  "The available values for {modality_phrase} indicate {numeric_finding} ({stat_phrase}),
although the evidence consists of only {sparse_observation_phrase}.",
]
LAB_RESULT_CAPTION_TEMPLATES = [
  "{modality_phrase}:  {numeric_finding}; {stat_phrase}.",
  "{numeric_finding} appears in recent {modality_phrase} results ({stat_phrase}).",
  "Recent lab evidence from {modality_phrase} shows {numeric_finding}; {stat_phrase}.",
  "The resulted {modality_phrase} panel contributes {numeric_finding}:  {stat_phrase}.",
  "{stat_phrase} supports {numeric_finding} in the {modality_phrase} category.",
  "Lab context before the decision point includes {numeric_finding} from
{modality_phrase}; {stat_phrase}.",
  "{modality_phrase} result summary:  {numeric_finding}; {stat_phrase}.",
  "A recent {modality_phrase} finding is {numeric_finding}, recorded as {stat_phrase}.",
]
WAVEFORM_GENERIC_TEMPLATES = [
  "{modality_phrase}:  {pattern_phrase}; {stat_phrase}.",
  "The {modality_phrase} signal shows {pattern_phrase} ({stat_phrase}).",
  "{modality_phrase} provides waveform evidence of {pattern_phrase}; {stat_phrase}.",
  "For {modality_phrase}, the dominant waveform pattern is {pattern_phrase}
({stat_phrase}).",
]
WAVEFORM_ECG_II_TEMPLATES = [
  "ECG Lead II: {pattern_phrase}; {stat_phrase}.",
  "Lead II waveform evidence includes {pattern_phrase} ({stat_phrase}).",
  "The cardiac waveform channel shows {pattern_phrase}; {stat_phrase}.",
  "ECG channel summary:  {pattern_phrase}; {stat_phrase}.",
]
WAVEFORM_PLETH_TEMPLATES = [
  "Pleth waveform:  {pattern_phrase}; {stat_phrase}.",
  "The pulse-ox waveform shows {pattern_phrase} ({stat_phrase}).",
  "Peripheral pulse waveform summary:  {pattern_phrase}; {stat_phrase}.",
  "Pleth channel evidence is summarized as {pattern_phrase}, with {stat_phrase}.",
]
```

```
WAVEFORM_RESP_TEMPLATES = [
  "Resp waveform:  {pattern_phrase}; {stat_phrase}.",
  "The respiration waveform shows {pattern_phrase} ({stat_phrase}).",
  "Breathing waveform evidence includes {pattern_phrase}; {stat_phrase}.",
  "Respiration waveform summary:  {pattern_phrase}; {stat_phrase}.",
]
CGM_WAVEFORM_CAPTION_TEMPLATES = {
  "glucose_summary":  [
    "Median sensor glucose over {period} was {median} mg/dL, {band_phrase}; readings
{lo}-{hi} mg/dL.",
    "Over {period}, median glucose was {median} mg/dL, {band_phrase}, with readings from
{lo} to {hi} mg/dL.",
  ],
  "glucose_trend":  [
    "Glucose was {direction} over {period}, with a fitted slope of {slope} mg/dL per
hour.",
    "The fitted glucose slope over {period} was {slope} mg/dL per hour; glucose was
{direction}.",
  ],
  "glucose_last":  [
    "The last reading before the decision was {last} mg/dL, {m} minutes before it.",
    "The most recent glucose, {m} minutes before the decision, was {last} mg/dL.",
  ],
  "time_in_ranges":  [
    "Time in ranges over {period}:  {p_lt54}% below 54, {p_lt70}% below 70, {p_70_180}% at
70-180, {p_gt180}% above 180 and {p_gt250}% above 250 mg/dL.",
    "Of valid readings in {period}, {p_70_180}% were at 70-180 mg/dL, {p_lt70}% below 70
({p_lt54}% below 54) and {p_gt180}% above 180 ({p_gt250}% above 250).",
  ],
  "variability":  [
    "The p10-p90 glucose spread over {period} was {spread} mg/dL, at least 60 mg/dL.",
    "Glucose variability in {period}:  p10-p90 spread {spread} mg/dL, 60 mg/dL or more.",
  ],
  "no_glucose":  [
    "No usable sensor glucose was available in {period}.",
    "Sensor glucose was unavailable for {period}.",
  ],
}
CGM_NUMERIC_CAPTION_TEMPLATES = {
  "basal_zero_throughout":  [
    "No basal insulin was delivered in {period}.",
    "Basal delivery was zero throughout {period}.",
  ],
  "basal_steady":  [
    "Basal insulin ran at a constant {rate} U/h over {period}, {total} U in total.",
    "Over {period}, basal delivery held at {rate} U/h, totalling {total} U.",
  ],
  "basal_varying":  [
    "Basal insulin over {period} averaged {mean} U/h, ranging {min}-{max} U/h, {total} U
in total.",
    "Over {period}, basal delivery ranged {min}-{max} U/h (mean {mean} U/h), totalling
{total} U.",
  ],
  "iob":  [
    "Insulin on board before the decision was {iob} U.",
    "Just before the decision, insulin on board was {iob} U.",
  ],
  "cob":  [
    "Carbohydrate on board before the decision was {cob} g.",
    "Just before the decision, carbohydrate on board was {cob} g.",
  ],
}
```

RELATIONAL

```
TEMPORAL_CAPTION_TEMPLATES = [
  "Temporal pattern: {modality_phrase} is {trend_phrase} across {window_phrase};
{numeric_finding}.",
  "Over the monitored interval, {modality_phrase} is {trend_phrase} with
{numeric_finding}.",
  "The observed trajectory for {modality_phrase} is {trend_phrase}; {stat_phrase}.",
  "Near the decision point, {modality_phrase} shows {numeric_finding} and is
{trend_phrase}.",
]
CROSS_SIGNAL_CAPTION_TEMPLATES = [
  "{numeric_finding} and {waveform_finding} are directionally consistent.",
  "{numeric_finding} is {connector_evidence} {waveform_finding}.",
  "{waveform_finding} provides waveform support for {numeric_finding}.",
  "{numeric_finding} is not clearly matched by waveform evidence because
{uncertainty_phrase}.",
  "{cross_signal_finding}: {numeric_finding} {connector_additive} {waveform_finding}.",
  "Together, {numeric_finding} and {waveform_finding} support {clinical_state}.",
  "{cross_signal_finding} is present, based only on available numeric and waveform
evidence.",
  "{waveform_finding} and {numeric_finding} point to {clinical_state}, with proxy-level
certainty.",
  "{numeric_finding} contrasts with {waveform_finding}; signal interpretation is
{uncertainty_phrase}.",
  "Cross-signal evidence: {cross_signal_finding}; {supporting_findings}.",
]
CROSS_SIGNAL_EXPLICIT_FACT_TEMPLATES = [
  "{cross_signal_finding} is supported by the paired numeric and waveform observations.",
  "Matched numeric and waveform channels show {cross_signal_finding}.",
  "{cross_signal_finding} is present across the corresponding signal channels.",
]
CGM_TEMPORAL_CAPTION_TEMPLATES = {
  "bolus_event": [
    "A {dose} U bolus was recorded {m} minutes before the decision.",
    "Bolus: {dose} U, {m} minutes before the decision.",
  ],
  "carb_event": [
    "{grams} g of carbohydrate was recorded {m} minutes before the decision.",
    "Carbohydrate: {grams} g, {m} minutes before the decision.",
  ],
  "small_deliveries_one": [
    "One run of deliveries under 0.25 U per row ({total} U) was not counted as a bolus.",
    "A run of sub-0.25 U deliveries totalling {total} U was not counted as a bolus.",
  ],
  "small_deliveries_many": [
    "{count} runs of deliveries under 0.25 U per row ({total} U) were not counted as
boluses.",
    "{count} runs of sub-0.25 U deliveries totalling {total} U were not counted as
boluses.",
  ],
  "basal_zero_ended": [
    "Basal delivery was zero for {dur} minutes, ending {end_m} minutes before the
decision.",
    "A {dur}-minute zero-basal stretch ended {end_m} minutes before the decision.",
  ],
  "basal_zero_ongoing": [
    "Basal delivery had been zero for {dur} minutes and still was in the slot before the
decision.",
    "A {dur}-minute zero-basal stretch ran up to the slot before the decision.",
  ],
  "basal_changes_one": [
```

```
    "The basal rate changed once, {m} minutes before the decision.",
    "One basal rate change, {m} minutes before the decision.",
  ],
  "basal_changes_many": [
    "The basal rate changed {n} times, last {m} minutes before the decision.",
    "{n} basal rate changes, the latest {m} minutes before the decision.",
  ],
  "temporal_empty": [
    "No bolus, carbohydrate entry, small delivery run or basal change was recorded in
{period}.",
    "{period_capitalized} has no bolus, carbohydrate entry, small delivery run or basal
change.",
  ],
}
```

GLOBAL

```
CONTEXT_CAPTION_TEMPLATES = [
  "Presentation context includes {supporting_findings}.",
  "Available contextual evidence includes {supporting_findings}.",
  "The available clinical context mentions {supporting_findings}.",
  "Before the decision point, the contextual record contains {supporting_findings}.",
  "Contextual cues before the decision point include {supporting_findings}.",
  "Relevant presentation and workflow context includes {supporting_findings}.",
  "Earlier clinical context includes {supporting_findings}.",
  "Pre-decision context includes {supporting_findings}.",
]
GLOBAL_CAPTION_TEMPLATES = [
  "Overall observed state before the decision: {clinical_state}, supported by
{supporting_findings}.",
  "The pre-decision state is most consistent with {clinical_state}; key evidence is
{supporting_findings}.",
  "Across directly observed context and signals, the dominant state is {clinical_state},
with {supporting_findings}.",
]
CGM_CONTEXT_CAPTION_TEMPLATES = {
  "context": [
    "Patient context: {items}.",
    "Patient: {items}.",
  ],
  "context_empty": [
    "No patient context was recorded.",
    "Patient context is not recorded.",
  ],
}
CGM_GLOBAL_CAPTION_TEMPLATES = {
  "global_states": [
    "Overall state: {states}.",
    "Overall pre-decision state: {states}.",
  ],
  "global_none": [
    "Overall state: no selected glycaemic state; direct pre-decision evidence was
limited.",
    "Overall state: no glycaemic state was selected; direct pre-decision evidence was
limited.",
  ],
}
```

ACTION

```
TREATMENT_ACTION_EVIDENCE_TEMPLATES = {
  "supported": [
    "Observed findings relevant to {action_evidence_domain_phrase} include
{action_evidence_clause}.",
    "The available pre-decision evidence for {action_evidence_domain_phrase} is
{action_evidence_clause}.",
  ],
  "secondary_only_evidence": [
    "The observed window includes {action_evidence_clause}; this is nonspecific supporting
evidence for {action_evidence_domain_phrase} rather than a complete indication.",
    "{action_evidence_clause} is relevant to {action_evidence_domain_phrase}, although the
available evidence is indirect or incomplete.",
  ],
  "fallback_visible_evidence": [
    "Visible pre-decision findings include {action_evidence_clause}; together they provide
indirect or mixed context for {action_evidence_domain_phrase}, not a complete
indication.",
    "The observed state contains {action_evidence_clause}; these facts are relevant
background for {action_evidence_domain_phrase}, although none is a direct domain-specific
match.",
  ],
  "label_review_context_only": [
    "Visible pre-decision findings include {action_evidence_clause}; they are retained as
clinical context, not as a direct indication for {action_evidence_domain_phrase}.",
    "The available state shows {action_evidence_clause}; because this recorded treatment
can represent context-specific care, these facts are described as background for
{action_evidence_domain_phrase} rather than a definitive rationale.",
  ],
  "label_review_no_visible_context": [
    "The recorded treatment belongs to a context-dependent category; no reliable
pre-decision fact is stated as a direct indication.",
    "This action category can reflect context-specific care, and the visible window does
not support a label-specific rationale statement.",
  ],
  "limited_observable_evidence": [
    "The available pre-decision record contains limited observable evidence specific to
this action domain.",
    "No compatible state fact was selected for an action-specific evidence statement in
this window.",
  ],
  "no_immediate_treatment_evidence": [
    "The observed pre-decision state does not assert a specific immediate treatment-level
action from the available evidence.",
    "No immediate treatment domain is selected from the visible pre-decision context and
signals.",
  ],
}
DIAGNOSTIC_ACTION_EVIDENCE_TEMPLATES = {
  "direct_note_indication": [
    "The pre-decision note documents this diagnostic indication:
{action_evidence_clause}.",
    "A documented indication before T0 was: {action_evidence_clause}.",
  ],
  "inferred_indication": [
    "The {action_evidence_domain_phrase} workup was likely prompted by
{action_evidence_clause}; this is an inferred indication from pre-T0 evidence, not an
explicitly documented reason.",
    "Pre-T0 findings suggest a possible indication for {action_evidence_domain_phrase}:
{action_evidence_clause}; the note did not state this rationale directly.",
  ],
```

```
  "inferred_general_indication":  [
    "The {action_evidence_domain_phrase} workup was likely part of broader diagnostic
assessment, but no specific pre-TO indication was available.",
    "A diagnostic indication for {action_evidence_domain_phrase} is plausible from the
clinical setting, although the available record does not identify a specific reason.",
  ],
  "supported":  [
    "Diagnostic-workup evidence for {action_evidence_domain_phrase} includes
{action_evidence_clause}.",
    "The pre-decision record supports {action_evidence_domain_phrase} through
{action_evidence_clause}.",
  ],
  "secondary_only_evidence":  [
    "The observed window includes {action_evidence_clause}; this is indirect supporting
context for {action_evidence_domain_phrase}.",
    "Available context such as {action_evidence_clause} is relevant to
{action_evidence_domain_phrase}, but is not a complete indication.",
  ],
  "fallback_visible_evidence":  [
    "Visible pre-decision findings include {action_evidence_clause}; they provide broad
context for {action_evidence_domain_phrase}, not a definitive indication.",
    "The available state shows {action_evidence_clause}; its relationship to
{action_evidence_domain_phrase} remains indirect or mixed.",
  ],
  "label_review_context_only":  [
    "Visible pre-decision findings include {action_evidence_clause}; they are clinical
context for {action_evidence_domain_phrase}, not a label-specific reason.",
    "The available state shows {action_evidence_clause}; the diagnostic action is retained
without inferring an exact indication.",
  ],
  "limited_observable_evidence":  [
    "A diagnostic action was recorded, but no specific pre-decision indication was
identifiable from the available context and signals.",
    "The visible window provides limited observable evidence for this diagnostic workup.",
  ],
  "no_immediate_treatment_evidence":  [
    "No immediate treatment action is asserted; the observed decision context is
diagnostic or otherwise non-treatment focused.",
    "The visible state does not establish an immediate treatment response, while
diagnostic context may still be present.",
  ],
}
OR_ACTION_EVIDENCE_TEMPLATES = {
  "supported":  [
    "Observed findings relevant to {action_evidence_domain_phrase} include
{action_evidence_clause}.",
    "The available pre-decision evidence for {action_evidence_domain_phrase} is
{action_evidence_clause}.",
    "For the current decision, visible evidence linked to {action_evidence_domain_phrase}
includes {action_evidence_clause}.",
    "Evidence aligned with {action_evidence_domain_phrase} includes
{action_evidence_clause}.",
    "The available state includes {action_evidence_clause}; these observations are
relevant to {action_evidence_domain_phrase}.",
    "Findings that bear on {action_evidence_domain_phrase} are {action_evidence_clause}.",
  ],
  "no_immediate_treatment_evidence":  [
    "No immediate action domain is selected from the visible pre-decision context and
signals.",
    "The caption records the current state without assigning an immediate action domain.",
    "Current findings are described without linking them to a recorded immediate action.",
    "No action-specific rationale is asserted from the observed pre-decision state.",
```

```
    "The available pre-T0 state does not select a specific immediate response.",
  ],
}
CGM_ACTION_EVIDENCE_TEMPLATES = {
  "evidence_for": [
    "Findings for an insulin bolus: {findings}.",
    "Evidence for a bolus: {findings}.",
  ],
  "evidence_against": [
    "Findings against an insulin bolus: {findings}.",
    "Evidence against a bolus: {findings}.",
  ],
  "evidence_none": [
    "No finding for or against an insulin bolus was selected from the observed window.",
    "No finding for or against a bolus was selected.",
  ],
}
```