LENS: LLM-Enabled Narrative Synthesis for Mental Health by Aligning Multimodal Sensing with Language Models

Wenxuan Xu[*1], **Arvind Pillai**[*1], **Subigya Nepal**[2], **Amanda C Collins**[3, 4], **Daniel M Mackin**[1], **Michael V Heinz**[1], **Tess Z Griffin**[1], **Nicholas C Jacobson**[1], **Andrew Campbell** [1],

[1]Dartmouth College, [2]University of Virginia,
[3]Massachusetts General Hospital, [4]Harvard Medical School,
[*]Equal contribution,

Correspondence: {wenxuan.xu.gr, arvind.pillai.gr}@dartmouth.edu

Abstract

Multimodal health sensing offers rich behavioral signals for assessing mental health, yet translating these numerical time-series measurements into natural language remains challenging. Current LLMs cannot natively ingest long-duration sensor streams, and paired sensor–text datasets are scarce. To address these challenges, we introduce LENS, a framework that aligns multimodal sensing data with language models to generate clinically grounded mental-health narratives. LENS first constructs a large-scale dataset by transforming Ecological Momentary Assessment (EMA) responses related to depression and anxiety symptoms into natural-language descriptions, yielding over 100,000 sensor–text QA pairs from 258 participants. To enable native time-series integration, we train a patch-level encoder that projects raw sensor signals directly into an LLM's representation space. Our results show that LENS outperforms strong baselines on standard NLP metrics and task-specific measures of symptom-severity accuracy. A user study with 13 mental-health professionals further indicates that LENS-produced narratives are comprehensive and clinically meaningful. Ultimately, our approach advances LLMs as interfaces for health sensing, providing a scalable path toward models that can reason over raw behavioral signals and support downstream clinical decision-making [1].

1 Introduction

Mental health conditions on the spectrum of anxiety and depression affect an estimated 18% and 9.5% of adults in the United States each year (Johns Hopkins Medicine, 2023). Traditional screening methods typically rely on structured clinical interviews and validated self-report instruments such as the Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2001) and the Generalized Anxiety Disorder scale (GAD-2) (Spitzer et al., 2006). However, these assessments are limited by their high burden on clinicians, dependence on retrospective self-reports, and reduced ecological validity because they are administered in controlled settings that do not capture an individual's real-world context (Abd-Alrazaq et al., 2023).

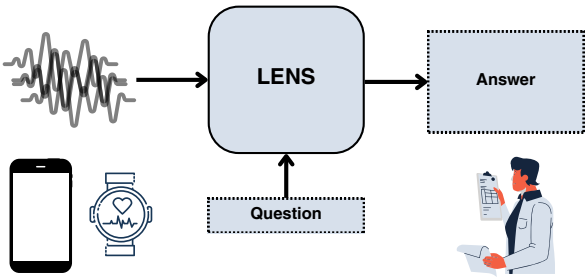

Figure 1: **Illustration of the LENS idea.** Given sensor signals from mobile and wearable devices, LENS takes a question as input and produces a natural-language description. Clinicians can then view an interpretable snapshot of the user's mental state instead of raw sensor streams.

To address these challenges, recent work has used mobile and wearable technologies to collect passive sensing data (for example, phone usage, speech features, and heart rate) and to administer ecological momentary assessments (EMA) (Xu et al., 2023; Gomes et al., 2023; Nepal et al., 2024a,b). A growing body of evidence shows that behavioral and physiological signals such as activity levels, sleep patterns, mobility, voice characteristics, and smartphone interactions can indicate the severity of depression and anxiety symptoms (Wang et al., 2014; Sheikh et al., 2021; Jacobson et al., 2021; Pillai et al., 2023). Together, these findings highlight passive sensing as a promising complementary approach for mental health monitoring.

In parallel, recent work has demonstrated the potential of large language models (LLMs) for mental health assessment. Studies show that prompt engi-

[1]Code available at https://github.com/Wen-xuan-Xu/LENS

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17458–17481
July 2-7, 2026

neering and fine-tuning can enable depression detection, symptom severity inference, physiological indicator prediction, and even generation of psychological rationales (Yang et al., 2024; Moon et al., 2025; Kim et al., 2024). Despite these promising capabilities, LLMs struggle with long-duration time series due to limitations in context length and tokenizer design, which prevent them from directly ingesting raw numerical sequences (Spathis and Kawsar, 2024). In contrast, multimodal vision and language models benefit from mature frameworks that support native visual understanding (Li et al., 2022; Dosovitskiy et al., 2021; Radford et al., 2021). However, comparable methods for native time-series integration remain scarce (Tan et al., 2025). As a result, the few studies that combine passive health sensing with LLMs rarely operate directly on raw sensor streams (Kim et al., 2024; Justin et al., 2024; Englhardt et al., 2024). Overall, progress in integrating behavioral sensing data with language is limited by the lack of large datasets that pair raw sensor streams with text and by the limited capabilities of existing methods that align time-series signals with language models.

We envision LENS as a *pre-consultation monitoring tool* for individuals already receiving clinical care for depression or anxiety, rather than a population-level screening or diagnostic instrument. In this intended workflow, clinicians receive a naturalistic behavioral summary of a patient's recent state derived from passive sensing, enabling more informed and time-efficient consultations 1. Toward this goal, our contributions are threefold. **(1) LENS data synthesis pipeline:** We introduce a pipeline that generates semantic mental health descriptions of multimodal sensing data from EMA responses (Section 3.1), producing a dataset of more than 100,000 sensor–text pairs and directly addressing the shortage of such resources. **(2) LENS training:** We propose a training strategy based on a patch-level time-series encoder that projects sensor signals into the language model's representation space (Section 3.2). By interleaving time-series and text embeddings and using a two-stage curriculum, we show that LENS can generate clinically grounded narratives. **(3) Comprehensive evaluation:** We evaluate our approach on a clinical dataset of 258 participants comprising 50,957 unique samples (Section 4.2). Beyond custom LLM metrics, we conduct a user study with 13 mental health experts who manually assessed 117 narratives.

2 Related Work

Recent work explores mobile and wearable data for mental health prediction, primarily focusing on symptom classification (Kim et al., 2024; Englhardt et al., 2024) or improving reasoning via multimodal encoders (Justin et al., 2024). These systems prioritize prediction over natural language generation, often producing text as secondary reasoning traces. Furthermore, methods that serialize numerical data into tokens face significant scalability and tokenization constraints (Pillai et al., 2025; Spathis and Kawsar, 2024; Yoon et al., 2024), and deploying models on heterogeneous edge devices introduces additional challenges around data distribution and communication efficiency (Liu et al., 2025b). In contrast, LENS anchors narrative generation in raw sensor measurements and clinically validated PHQ/GAD items to produce meaningful symptom descriptions.

LLMs are increasingly used to synthesize paired datasets for time-series tasks by generating artificial sequences or surrogate signals for QA pairs and explanations (Xie et al., 2025; Li et al., 2025b; Yan et al., 2023; Imran et al., 2024). While scalable, these pipelines rely on synthetic inputs that fail to capture the complexity of real physiological and behavioral signals; long-tail distributional skew in generated outputs further complicates quality assurance (Zhou et al., 2025). LENS instead pairs real-world sensor streams with clinical assessments, using LLMs only to refine linguistic fluency via rigorous quality checks.

While multimodal alignment has progressed rapidly across vision–language understanding (Liu et al., 2023), embodied view synthesis (Wang et al., 2025), and video-based physical reasoning (Liu et al., 2026), comparable integration for time-series data remains limited. Existing text-based serialization is constrained by numerical encoding and context length (Xue and Salim, 2023; Gruver et al., 2023), while vision-based methods convert series into images, introducing plot-engineering biases and indirect representations (Yoon et al., 2024; Zhang et al., 2023). Alignment-based approaches project encoder representations into LLM hidden states (Ming et al., 2023; Xie et al., 2025). However, by optimizing for synthetic QA, they rarely generate natural language aligned with real-world clinical constructs, and the reliability of standard benchmarks in specialized domains remains an open concern (Yao et al., 2025). Our work dif-

fers from these approaches by producing symptom-oriented narratives tied to psychometric instruments and grounded directly in raw signals

3 Methods

LENS consists of two components: a scalable dataset-construction pipeline that transforms EMA responses into high-quality sensor–text pairs (Section 3.1), and a sensor–text alignment method that enables native integration of raw time-series signals into an LLM through a patch-based encoder (Section 3.2).

3.1 LENS: Dataset Construction

3.1.1 Study

Our longitudinal study investigates intra-day fluctuations in mental health symptoms among individuals diagnosed with major depressive disorder. In this 90-day study, we recruited participants aged 18 years and older, residing in the United States. Each participant wore a Garmin vivoactive 3 device and installed our Android application. This setup enabled the collection of passive sensing data, which are behavioral and physiological signals automatically recorded from smartphones and wearables, and ecological momentary assessments (EMAs), in which participants actively reported depression and anxiety symptoms experienced over the past four hours. Non-identifiable study data will be available at the NIMH Data Archive (NDA)[2]. Researchers requesting access must submit a Data Use Certification (DUC) through the NDA platform, which requires institutional sponsorship and the signature of an Authorized Institutional Business Official at an institution with an active Federal Wide Assurance.

The EMA consists of 14 items adapted from the PHQ-9 and GAD-4 (Table 2; Appendix §B). It is administered three times per day in the morning, afternoon, and evening, customized to each participant's waking time. Each item is rated on a continuous 0 ("Not at all") to 100 ("Constantly") scale. In addition to the EMAs, we collect mobile and wearable time-series signals, including GPS traces, step counts, accelerometer-derived zero-crossing rate (ZCR) and energy, conversation time, phone lock and unlock events, heart rate, sleep estimates, and stress levels. These signals have been shown to correlate with depressive and anxiety symptoms (Choudhary et al., 2022). To align self-reported EMAs with corresponding sensing data, we use each EMA's completion time to retrieve the preceding four hours of sensor data. This procedure results in a temporally aligned, multimodal dataset suitable for narrative synthesis and modeling. Ultimately, we utilize 50,957 EMAs from 258 participants. Although all participants carry an MDD diagnosis, the ordinal EMA distributions span the full 0–3 severity range across all nine PHQ-9 items, with over 42% of samples falling into the "No/Minimal" or "Mild" categories, ensuring sufficient within-cohort variability (Table 10).

3.1.2 Sensor Data Processing

The sensing data includes two types of signals: (1) continuous streams, which contain time-series data such as steps, heart rate, accelerometer-derived zero-crossing rate (ZCR) and energy, phone lock and unlock state, stress level, and GPS traces, and (2) aggregated streams, which contain daily or window-level values such as sleep duration from the previous night and total conversation time. After preprocessing, all signals are standardized to fixed sampling rates to ensure consistent temporal alignment across modalities (see Appendix §C for specific details).

3.1.3 LENS Narrative Synthesis

To address the lack of sensor-text datasets describing mental health symptoms, LENS constructs high-quality QA datasets from self-reported EMA responses to serve as ground truth for model training and evaluation. Because each numeric EMA response corresponds to a symptom intensity category aligned with DSM-5 semantics, these values can be converted into natural language descriptions of participants' experiences. We produce two QA datasets: (1) **Item-level QA dataset**: Each sample represents a single EMA item and supports question-specific supervision, and (2) **Summary-level QA dataset:** Each sample reflects all EMA questions and targets full-summary generation.

Responses → Answers (A). EMA responses are converted into narrative labels as shown in Figure 2. For each question, we first transform the EMA item into a symptom-focused template sentence. We then insert the frequency phrase associated with its numeric range (0–25: not at all, 26–50: sometimes, 51–75: often, 76–100: constantly) to obtain the raw item-level narrative. Question 14 is treated as binary, and overall severity is categorized as mild, moderate, or severe. Summary-level narratives

[2] https://nda.nih.gov

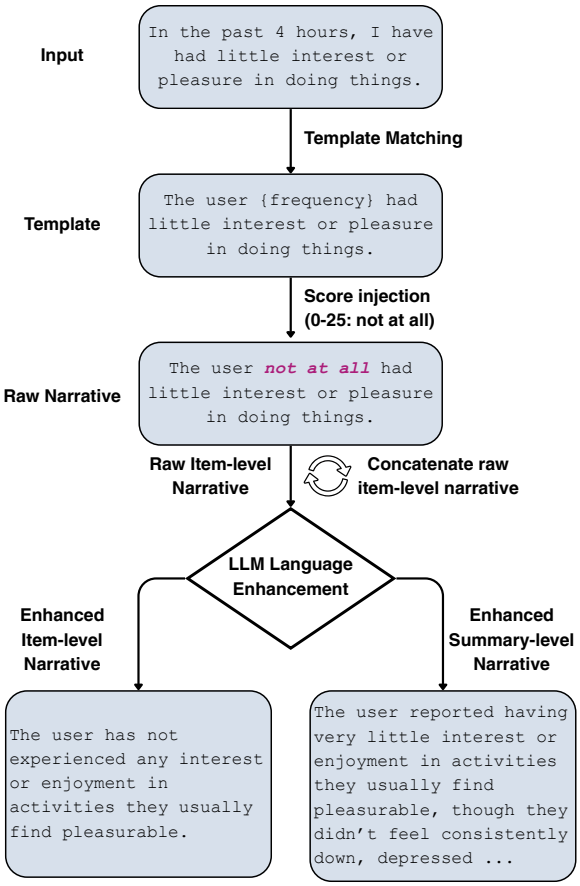

Figure 2: **Overview of the narrative synthesis pipeline.** EMA questions and responses are mapped to templates populated with corresponding frequency phrases. Subsequently, an LLM refines the text at both the item and summary levels (concatenated narratives) to enhance fluency and lexical diversity.

are formed by concatenating item-level narratives before applying LLM refinement. To reduce the mechanical tone of the templates, we use prompt-based rewriting with the locally deployed Qwen2.5-14B model. The system prompt enforces factual accuracy, consistent terminology, and stigma-free language, while the user prompt supplies the rule-based text and requests a fluent rewrite (See prompt template in Appendix §G). This process produces 101,914 item-level narratives and 50,957 summary-level narratives that serve as ground truth for training LENS.

Questions (Q). The ground-truth questions consist of item-level prompts for each EMA question and a summary-level prompt that covers all items along with the overall severity statement. To increase linguistic diversity, we use GPT-4o to generate paraphrased variants of every question. For each original prompt, we created ten semantically equivalent but lexically distinct phrasings. During QA dataset construction, one paraphrased variant is randomly sampled for each instance.

Quality Control. To ensure narrative reliability, we implement an automatic quality-control pipeline using a multi-model LLM-judge system (Figure 3). Each judge model follows an LLM-as-a-Judge prompting template (Li et al., 2025a) with rule-augmented instructions that embed evaluation principles, baseline references, and dimension-specific rubrics directly in the prompt. Judges compare the template-based and enhanced narratives in a pairwise format and output five dimension scores (1–5), confidence values, and a short rationale, providing structured and transparent assessments. Because individual LLM judges can be biased, we use three independent models (Mistral-7B (Jiang et al., 2023), Llama-3.1-8B (Grattafiori et al., 2024), and Qwen2.5-7B (Qwen et al., 2025)). Dimension scores are averaged and rounded, and their sum forms a total quality score; confidence values are averaged similarly. A narrative is accepted only if its average total score exceeds 20 (out of 25) and its mean confidence exceeds 0.8. Otherwise, it is returned for regeneration and re-evaluation (FAIL in Figure 3). This iterative refine-and-judge loop, akin to Refine-n-Judge (Cayir et al., 2025), filters out low-quality outputs and ensures that finalized narratives faithfully reflect EMA responses while improving fluency and diversity without introducing distortions. Importantly, the multi-stage LLM pipeline is used strictly for offline dataset construction and is never invoked at inference time; the deployed model consists solely of the patch-based encoder and the LLM backbone. Using commercial API pricing as a conservative upper bound, the total equivalent cost of constructing the full dataset is approximately $7.75 ($0.00015 per sample), ensuring high scalability (Table 9).

3.2 LENS: Training & Architecture

3.2.1 Time-series Encoder

We encode each time–series stream independently using a lightweight patch-based module that converts raw scalar values into language-model-compatible embeddings. Given a univariate sequence $S = \{s_t\}_{t=1}^{T}$, we first apply a reversible value normalization to stabilize scale while keeping absolute magnitudes recoverable. Let μ and σ denote the per-stream mean and standard deviation, then $\tilde{s}_t = \frac{s_t - \mu}{\sigma}$, and auxiliary statistics such as $(\mu, \sigma, m_{\min}, m_{\max})$ are inserted into the textual

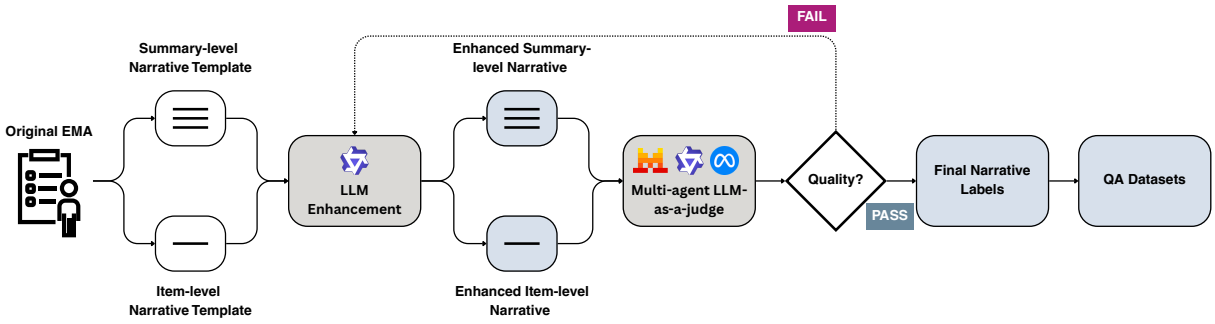

Figure 3: **LENS dataset construction pipeline.** EMA responses are first converted into item-level and summary template narratives, which are then rewritten into fluent, enhanced narratives using Qwen2.5-14B. A multi-agent LLM-as-a-judge system conducts automatic quality control, routing failed cases back for regeneration until they satisfy all criteria. The final accepted narratives are then paired with paraphrased question variants to construct the QA datasets.

prompt as metadata, allowing the model to reason about original numerical ranges inside the language space (Langer et al., 2025; Xie et al., 2025). The normalized sequence is divided into $N = \lceil T/w \rceil$ non-overlapping patches of width w, where $w = 8$ for all streams in our experiments. Each timestep is augmented with a learnable positional embedding $q_t \in \mathbb{R}^{d_p}$ with $d_p = 16$. For each patch, scalar values and positional codes are concatenated as follows:

$$u_i = [\tilde{s}_t; q_t]_{t=(i-1)w+1}^{iw} \in \mathbb{R}^{w(1+d_p)}$$

Each patch representation is projected into the hidden space of the pretrained language model via a multilayer perceptron f^{ts} consisting of 5 layers with hidden width 5120, matching the LLM embedding dimension d. The encoder outputs one embedding per patch,

$$z_i = f^{\text{ts}}(u_i) \in \mathbb{R}^d, \quad Z = \{z_1, \ldots, z_N\}$$

This design captures localized temporal patterns at the patch level while preserving absolute numerical semantics through reversible normalization.

3.2.2 Architecture

Given an instruction text $X = \{x_1, \ldots, x_M\}$ and K time–series streams $\{S^{(k)}\}_{k=1}^K$, we inject the normalization metadata described above into the text. After this step, the prompt $\tilde{X}$ contains special placeholder tokens `<ts>` and `</ts>` to mark the position of each stream. The text $\tilde{X}$ is then tokenized and passed through the pretrained LLM embedding layer f^{text}, producing a sequence of text embeddings $E_{\text{text}} = \{e_1, \ldots, e_L\} \in \mathbb{R}^{L \times d}$ (Figure 4). In parallel, each time–series stream $S^{(k)}$ is processed by the encoder, yielding patch embeddings $Z^{(k)} = \{z_1^{(k)}, \ldots, z_{N_k}^{(k)}\}$ with $z_i^{(k)} \in \mathbb{R}^d$. The multimodal embedding sequence is formed by concatenating the text embeddings and the patch embeddings at the positions referenced by the placeholders, i.e., $H = \text{concat}\left(E_{\text{text}}, Z^{(1)}, \ldots, Z^{(K)}\right) \in \mathbb{R}^{L_{\text{mm}} \times d}$, which interleaves natural–language and time–series representations in a single ordered context. This unified sequence is then fed into the subsequent transformer blocks of the pretrained LLM, enabling multimodal reasoning over both textual instructions and temporal patterns.

3.2.3 Training

Stage 1: Encoder Alignment. The first stage aims to establish a stable alignment between temporal features and their textual descriptions. We use the alignment dataset from ChatTS (Xie et al., 2025), which takes a QA form and probes time–series understanding across different signal types and temporal behaviors, enabling the encoder to capture trends, correlations, and local pattern variations rather than raw numeric fluctuations. However, our final objective is to generate clinically coherent symptom narratives rather than attribute-only descriptions. To prevent the encoder from overspecializing on low-level attribute queries, we interleave a small portion of narrative and general QA samples into the alignment corpus (8:1:1 with alignment data), giving the model early exposure to natural question forms and narrative structure and better preparing it for downstream symptom-focused generation.

Stage 2: Supervised Fine-Tuning on Symptom Narratives. After encoder alignment, we train the model to perform symptom-centered question answering and narrative generation. We fine-tune

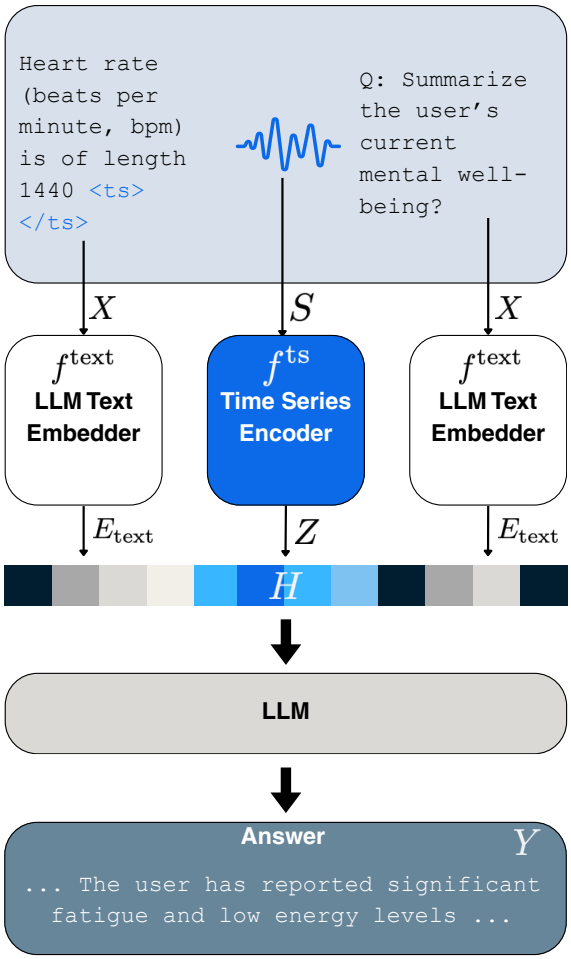

Figure 4: **LENS Architecture.** The model accepts multimodal inputs consisting of description text (e.g., "Heart rate..."), instruction text (e.g., "Summarize the user's current mental well-being?") and raw time-series sensor streams (e.g., heart rate). The text is processed by a pretrained LLM text embedder (f^{text}), while time-series data is encoded by a trainable patch-based encoder (f^{ts}). The resulting embeddings are concatenated into a unified sequence (H) and processed by the LLM backbone to generate a natural language response (Y).

LENS on the two EMA-derived QA datasets described in Section 3.1.3, which respectively provide item-level supervision for single-symptom interpretation and summary-level supervision for multi-symptom synthesis. To prevent the model from overfitting to a single response style and to enhance its ability to follow structured outputs, we additionally include an instruction-following (IF) dataset constructed from predefined templates. This dataset exposes the model to consistent answer formats, encouraging stable response organization under different prompts. Because real-world time-series queries vary widely in temporal resolution, we further interleave an alignment-random subset in which sequence lengths are uniformly sampled between 64 and 1024, enabling the model to generalize across heterogeneous sampling rates and time spans. During supervised fine-tuning, the four datasets are mixed using a 0.3:0.3:0.2:0.2 ratio (item-level QA : summary-level QA : IF : alignment-random), which balances symptom narrative grounding, structural instruction patterns, and robustness to sequence-length variability.

Given an instruction text $X = \{x_1, \ldots, x_M\}$ and K time–series streams $\{S^{(k)}\}_{k=1}^{K}$, the model is trained in a standard supervised autoregressive manner. For a target response $Y = \{y_1, \ldots, y_U\}$, the objective is

$$\mathcal{L} = -\sum_{t=1}^{U} \log p_\phi\big(y_t \mid X, \{S^{(k)}\}_{k=1}^{K}, y_{1:t-1}\big)$$

where ϕ denotes all trainable parameters. We adopt full-parameter fine-tuning, updating the time-series encoder, pretrained LLM, and the text embedder. See Appendix §A for hardware and batching details.

4 Experiments

4.1 Evaluation Settings

Baselines. To contextualize LENS's performance, we compare it with baselines that follow common strategies for modeling time-series signals with LLMs. Prior work often encodes time-series values as raw text (Kim et al., 2024; Gruver et al., 2023), so we adopt the same principles and utilize the Qwen2.5-14B (Yang et al., 2025) with few-shot prompting as the backbone; we refer to this as **TS-Text**. Another promising approach converts time-series data into visual representations to support downstream reasoning (Yoon et al., 2024; Liu et al., 2025a). Following this approach, we transform each signal into a plot and generate narratives using the Qwen-2.5VL-32B model in a few-shot setting. This baseline is denoted as **TS-Image**. Inference for all baseline models utilizes default configurations to ensure a controlled comparison, and we utilize participant-level splits (approx. 70:15:15) to prevent information leakage, ensuring that all data from a given individual appears in only one split (Appendix §A).

Metrics. We evaluate model performance using three categories of metrics. **(1) Linguistic Metrics**: Generated narratives are assessed with ROUGE-1/2/L (Lin, 2004), BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and

Lavie, 2005), and BERTScore (Zhang et al., 2020), which measure lexical overlap, structural consistency, and semantic similarity with reference narratives. **(2) Symptom-grounded Evaluation**: To assess clinical alignment, we employ a structured LLM-as-a-judge protocol using `gpt-4.1-mini` (See prompt template in Appendix §G. The judge is queried with temperature set to 0 to ensure deterministic and reproducible evaluations). For each of the PHQ-related symptom categories, the judge first outputs a JSON record containing `ref_presence`, `ref_severity`, `pred_presence`, and `pred_severity`, indicating whether the model omits or hallucinates symptom dimensions and whether predicted severity is faithful to the ordinal reference. **(3) Item-level QA Evaluation**: For single-question outputs, evaluation reduces to a record specifying `ref_severity` and `pred_severity`; symptom presence is implicit in the question itself, allowing severity correctness to be measured without aggregating across narrative spans. Detailed definitions of coverage, presence-aware severity alignment, and the weighting procedure used for ordinal scoring are provided in Appendix §D.

User study. To assess the characteristics of LENS-generated narratives, we recruited 13 mental health experts to evaluate LENS (14B; zero-shot) along with the two baselines: TS-Image (32B, few-shot) and TS-Text (14B, few-shot), across four rating dimensions. Each example is a narrative paired with a table of 12 ground-truth symptoms and assessed as follows: (1) Comprehensiveness: "Does the narrative mention the symptom?"; (2) Accuracy: "Does the narrative accurately describe the symptom severity (including synonyms)?"; (3) Clinical Utility: "Does the narrative provide clinically useful information that informs care?"; and (4) Language Cohesion: "Is the narrative coherent, easy to understand, and focused on depression and anxiety symptoms?" Comprehensiveness and accuracy were collected as yes/no responses and grouped into five categories. For comprehensiveness: 0–2 = Very Poor, 3–5 = Poor, 6–7 = Adequate, 8–11 = Good, 12 = Excellent. For accuracy: 0–2 = Major Discrepancy, 3–5 = Substantial Discrepancy, 6–7 = Partial Agreement, 8–11 = High Agreement, 12 = Complete Agreement. Clinical utility and language cohesion were rated on a 1–5 Likert scale. In total, we evaluated 117 narratives, with each expert rating 9 narratives (3 samples from each model). Additional information regarding the user study and expert background is provided in Appendix §F. To compare the models across the four domains, we first computed per-rater mean scores for each model by averaging over the three evaluated samples, and then performed paired t-tests across raters with Bonferroni correction (Weisstein, 2004). Effect sizes are reported using Cohen's d_z (Goulet-Pelletier and Cousineau, 2018).

Method	Linguistic Metrics		LLM-as-a-Judge Metrics	
	ROUGE-L	BERTScore	Coverage	Presence Alignment
Summary-level Evaluation				
LENS	**0.409**	**0.775**	**0.801**	**0.601**
TS-Text	0.151	0.630	0.614	0.372
TS-Image	0.373	0.764	0.740	0.579
Item-level Evaluation				
LENS	**0.603**	**0.832**	–	**0.732**
TS-Text	0.142	0.604	–	0.665
TS-Image	0.136	0.602	–	0.687

Table 1: **Results on Summary and Item level generation.** Additional metrics are provided in Appendix §E.

4.2 Results

LENS consistently outperforms baselines in generating comprehensive clinical narratives. In the summary-level evaluation (Table 1), LENS achieves the highest scores across all linguistic metrics, recording a ROUGE-L of 0.409 and a BERTScore of 0.775. This represents a 9.6% and 1.4% improvement over the strongest baseline, TS-Image, which achieved 0.373 and 0.764, respectively. The significantly lower performance of TS-Text (ROUGE-L=0.151) highlights the difficulty of generating cohesive summaries from time-series text descriptions alone, whereas LENS effectively integrates multi-modal signals to produce structurally and semantically robust narratives.

LENS demonstrates superior grounding in clinical symptoms, minimizing hallucinations. From a clinical perspective, LENS achieves the highest alignment with ground-truth diagnoses. In summary-level generation, it attains a Symptom Coverage score of 0.801 and a Presence Alignment score of 0.601. This indicates that LENS is more capable of capturing the full spectrum of patient symptoms ($K = 14$ categories) without omitting critical indicators or fabricating non-existent ones, outperforming TS-Image (Coverage=0.740) and greatly surpassing TS-Text (Presence Alignment=0.372).

LENS exhibits strong performance in fine-grained, item-level query answering. Compared to summary-level narrative generation, the perfor-

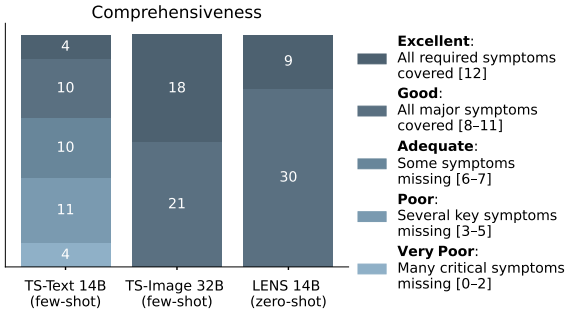

Figure 5: **User Study: Comprehensiveness.** Expert ratings compare how many symptoms each model's narrative successfully covers.

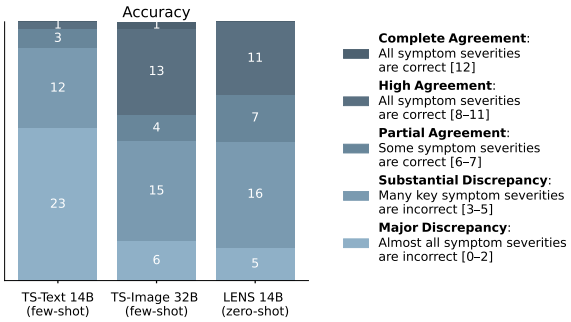

Figure 6: **User Study: Accuracy.** Expert ratings assess how well each model's narrative captures the severity of symptoms reported in the ground-truth EMA.

mance gap in the item-level evaluation is more pronounced, indicating that the task requires more precise retrieval (Table 1). For example, LENS achieves a ROUGE-L score of 0.603, which is more than four times higher than both TS-Text (0.142) and TS-Image (0.136). This trend holds for semantic and clinical metrics as well, with LENS securing the highest Presence Alignment of 0.732 (10.1% improvement over TS-Text and 6.6% over TS-Image) and a BERTScore of 0.832 (over 37% improvement against both baselines). These results suggest that while baselines struggle to isolate specific symptom intensities in a QA format, LENS maintains high fidelity to the reference severity levels.

Mental health experts rate LENS narratives significantly higher than TS-Text. As shown in Figures 5 to 7, LENS consistently outperforms TS-Text across all four evaluation dimensions. Specifically, LENS narratives receive significantly higher expert ratings for comprehensiveness (4.23 vs. 2.97; $T = 7.02$; $p < 0.01$; $d_z = 1.12$), accuracy (2.61 vs. 1.53; $T = 6.49$; $p < 0.01$; $d_z = 1.04$), clinical utility (3.25 vs. 1.84; $T = 7.16$; $p < 0.01$; $d_z = 1.15$), and language cohesion (3.82 vs. 2.23; $T = 7.34$; $p < 0.01$; $d_z = 1.17$). These results demonstrate that experts find narratives derived from native time-series integration to be not only more accurate but also more practically useful for care than those generated from text descriptions alone.

Expert evaluation demonstrates performance parity with larger visual baselines. Comparing LENS to TS-Image (Figures 5 to 7), we observe no statistically significant differences between the two models across any of the four evaluation dimensions. Expert ratings are comparable for comprehensiveness (4.23 vs. 4.46; $T = -2.47$; $p > 0.01$; $d_z = -0.39$), accuracy (2.61 vs. 2.69; $T = -0.42$; $p > 0.01$; $d_z = -0.06$), clinical utility (3.82 vs. 3.71; $T = 1.52$; $p > 0.01$; $d_z = 0.24$), and language cohesion (3.25 vs. 2.94; $T = 0.56$; $p > 0.01$; $d_z = 0.09$). These results indicate that LENS matches the performance of the 2.2× larger TS-Image while offering a significantly streamlined inference process. By replacing complex visual plot engineering with native time-series encoding, LENS achieves these results in a zero-shot setting. Interestingly, experts identified LENS as having better real-world applicability, evidenced by its marginally higher ratings for clinical utility and language cohesion.

Ablation and scalability analysis. We analyze LENS' architecture through ablation and scaling studies detailed in Appendix §E. An evaluation comparing LENS to a fine-tuned text-only baseline (TS-Text-FT) reveals that explicit time-series encoding is essential for capturing temporal structures; LENS significantly outperforms the text-only variant, particularly in item-level QA where the ROUGE-L gap is markedly wider (0.6030 vs. 0.1717) (Table 4). Furthermore, LENS demonstrates superior computational efficiency, requiring approximately 930 tokens per sample, a reduction of roughly 94% compared to verbose text serialization and 4× relative to vision-based models (Figure 8). We also observe that while basic instruction-following can be established with 10% of the training data, the full dataset is necessary to maximize clinical fidelity in complex narratives, with Presence Alignment improving as data size increases (Table 5). Ablations on encoder hyperparameters further confirm robustness: varying MLP depth ($L \in \{3, 5, 7\}$) and patch size ($p \in \{4, 8, 16\}$) yields fluctuations within 0.01 absolute across all metrics (Table 7). Finally, experiments with a lighter LENS-7B variant show that it maintains lin-

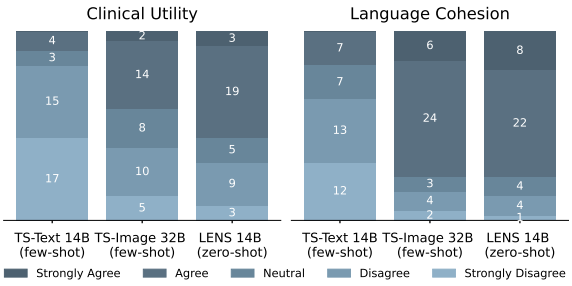

Figure 7: **User Study: Clinical Utility & Language Cohesion.** Comparing expert ratings for the usefulness of the narratives and the cohesiveness of the language.

guistic performance (ROUGE-L 0.409 vs. 0.410) while trailing slightly in complex clinical alignment (0.559 vs. 0.601) (Table 6). Additionally, a sensor mismatch experiment in which sensor streams are randomly shuffled across participants reveals that Presence Alignment drops by 20.1% while linguistic metrics remain unchanged (Table 8), providing direct evidence that LENS grounds clinical content in raw sensor signals rather than merely reproducing stylistic patterns from the training labels.

5 Conclusion

In this work, we introduced LENS, a framework for generating clinical narratives by aligning raw sensor streams with LLMs. Through a scalable data-synthesis pipeline and a patch-based time-series encoder, LENS addresses two key challenges: the scarcity of paired sensor–text data and the difficulty of processing long-horizon numerical signals within LLMs. Using a dataset of 50,957 samples, our evaluation showed that LENS produced linguistically fluent and clinically accurate descriptions of depression and anxiety symptoms. LENS also matched the performance of $2.2\times$ larger vision-based models while avoiding the overhead of plot generation, and mental-health experts rated its narratives as more applicable and clinically useful than text-based baselines. Future work should examine LENS's transferability to broader demographics and additional psychological instruments. Overall, LENS represents a fundamental step toward ecologically valid mental-health monitoring by transforming raw sensor data into interpretable clinical insights.

Limitations

Several limitations of this work warrant consideration. First, LENS currently focuses on depression and anxiety within a specific clinical population. While these conditions are well-suited for establishing a sensor-to-language framework, the findings may not yet generalize to other demographics, psychological instruments, or broader populations. Given the pipeline's modular design, future work should evaluate its extensibility to alternative disorders. Second, the framework analyzes behavior over four-hour windows. This captures immediate states but may miss longitudinal trends, such as relapse signals or treatment responses that unfold over months. Future extensions could employ hierarchical modeling or memory-augmented architectures to integrate these short-term representations into longer-horizon reasoning.

Third, this study utilized models from the Qwen family to maintain a controlled experimental environment rather than to optimize for peak performance. Consequently, we did not assess cross-model robustness. Future research should examine how different architectures, including instruction-tuned or domain-adapted variants, vary in their ability to interpret sensor data. Fourth, passive sensing lacks the situational context necessary to fully disambiguate certain signals. For instance, an elevated heart rate could indicate physical exertion, illness, or acute stress. Incorporating richer contextual data or structured user input remains a necessary step for reducing such ambiguity. Finally, the iterative Refine-n-Judge loop improves output quality but introduces computational overhead. This may challenge real-time deployment or large-scale application under constrained budgets. Exploring model distillation or more efficient training strategies could help mitigate these scaling issues.

Ethical Considerations

Study Participation. IRB-approved procedures involved 258 analyzed participants (from 300 recruited) who provided informed consent. Inclusion required a SCID-verified MDD diagnosis, excluding bipolar disorder, psychosis, or active suicidality. The sample was 84% female and 79% White (12% Hispanic/Latino) with a mean age of 40. Most participants were college-educated (93%) and employed (61%), with incomes aligning with national distributions. Compensation included $1 per completed EMA and a $50 bonus for 90% completion. Safety was maintained via automated suicidality alerts and continuous clinical oversight.

Systems & Software. Modeling was conducted using local LLMs within a secure, closed environment to prevent data transmission to external APIs or cloud services. Access was restricted to authorized, IRB-trained researchers via role-based controls and audit logging. Collectively, these safeguards reduced the risk of re-identification of personally identifiable information and ensured adherence to IRB-approved data protection and privacy protocols.

Adverse Usage. While LLMs offer significant potential for interpreting mental health time-series data, their application requires caution to protect participant safety and clinical integrity. These narratives must augment existing clinical workflows rather than serve as standalone diagnostic tools, as over-reliance on automated summaries may obscure critical nuances within the raw sensor data. Furthermore, the utilization of sensitive signals, such as GPS traces, necessitates strict governance to prevent unauthorized surveillance or secondary data use. Given that LLMs cannot entirely eliminate the risk of hallucinations, all generated outputs must be reviewed by qualified experts to ensure clinical judgment remains the final authority for interpretation and intervention.

Acknowledgments

The research presented in this paper was supported by the **National Institute of Mental Health (NIMH)** under award number **R01MH123482-01**, and by **Evergreen: A Generative AI and Behavioral Sensing Digital Ecosystem to Promote Student Wellness and Flourishing.** This work was further enabled by philanthropic gifts to Dartmouth College dedicated to advancing AI-supported well-being and student flourishing. The contents of this manuscript are solely the responsibility of the authors and do not necessarily represent the views of the funding agencies. The funders had no role in the study design, data collection, analysis, interpretation, or the preparation of this manuscript.

References

Alaa Abd-Alrazaq, Rawan AlSaad, Farag Shuweihdi, Arfan Ahmed, Sarah Aziz, and Javaid Sheikh. 2023. Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression. *NPJ Digital Medicine*, 6(1):84.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Derin Cayir, Renjie Tao, Rashi Rungta, Kai Sun, Sean Chen, Haidar Khan, Minseok Kim, Julia Reinspach, and Yue Liu. 2025. Refine-n-judge: Curating high-quality preference chains for llm-fine-tuning. *Preprint*, arXiv:2508.01543.

Soumya Choudhary, Nikita Thomas, Janine Ellenberger, Girish Srinivasan, and Roy Cohen. 2022. A machine learning approach for detecting digital behavioral patterns of depression using nonintrusive smartphone data (complementary path to patient health questionnaire-9 assessment): Prospective observational study. *JMIR Formative Research*, 6(5):e37736.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Zachary Englhardt, Chengqian Ma, Margaret E. Morris, Xuhai "Orson" Xu, Chun-Cheng Chang, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2024. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–25. ArXiv:2311.13063 [cs].

Nuno Gomes, Matilde Pato, Andre Ribeiro Lourenco, and Nuno Datia. 2023. A survey on wearable sensors for mental health monitoring. *Sensors*, 23(3):1330.

Jean-Christophe Goulet-Pelletier and Denis Cousineau. 2018. A review of effect sizes and their confidence intervals, part i: The cohen'sd family. *The Quantitative Methods for Psychology*, 14(4):242–265.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36:19622–19635.

Sheikh Asif Imran, Mohammad Nur Hossain Khan, Subrata Biswas, and Bashima Islam. 2024. LLaSA: A Multimodal LLM for Human Activity Analysis Through Wearable and Smartphone Sensors. *arXiv preprint*. ArXiv:2406.14498 [cs] version: 2.

Nicholas C Jacobson, Damien Lekkas, Raphael Huang, and Natalie Thomas. 2021. Deep learning paired

with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17–18 years. *Journal of affective disorders*, 282:104–111.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Johns Hopkins Medicine. 2023. Mental Health Disorder Statistics. https://www.hopkinsmedicine.org/health/wellness-and-prevention/mental-health-disorder-statistics. Accessed: November 14, 2025.

Cosentino Justin, Belyaeva Anastasiya, Liu Xin, Furlotte Nicholas, A., Yang Zhun, Lee Chace, Schenck Erik, Patel Yojan, Cui Jian, Schneider Logan, Douglas, Bryant Robby, Gomes Ryan, G., Jiang Allen, Lee Roy, Liu Yun, Perez Javier, Rogers Jameson, K., Speed Cathy, Tailor Shyam, and 15 others. 2024. Towards a personal health large language model.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-llm: Large language models for health prediction via wearable sensor data.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A. Xu, Winnie Chow, Martin Maritsch, 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. 2025. Opentslm: Time-series language models for reasoning over multivariate medical text- and time-series data. *Preprint*, arXiv:2510.02410.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025a. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. *arXiv preprint*. ArXiv:2411.16594 [cs].

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Preprint*, arXiv:2201.12086.

Zechen Li, Shohreh Deldari, Linyao Chen, Hao Xue, and Flora D. Salim. 2025b. SensorLLM: Human-Intuitive Alignment of Multivariate Sensor Data with LLMs for Activity Recognition. *arXiv preprint*. ArXiv:2410.10624 [cs].

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chunjiang Liu, Xiaoyuan Wang, Qingran Lin, Albert Xiao, Haoyu Chen, Shizheng Wen, Hao Zhang, Lu Qi, Ming-Hsuan Yang, Laszlo A. Jeni, Min Xu, and Yizhou Zhao. 2026. Mosiv: Multi-object system identification from videos. *Preprint*, arXiv:2603.06022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Haoxin Liu, Chenghao Liu, and B. Aditya Prakash. 2025a. A picture is worth a thousand numbers: Enabling LLMs reason about time series via visualization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7486–7518, Albuquerque, New Mexico. Association for Computational Linguistics.

Ji Liu, Beichen Ma, Qiaolin Yu, Ruoming Jin, Jingbo Zhou, Yang Zhou, Huaiyu Dai, Haixun Wang, Dejing Dou, and Patrick Valduriez. 2025b. Efficient federated learning with heterogeneous data and adaptive dropout. *ACM Trans. Knowl. Discov. Data*, 19(8).

Jin Ming, Wang Shiyu, Ma Lintao, Chu Zhixuan, Zhang James, Y., Shi Xiaoming, Chen Pin-Yu, Liang Yuxuan, Li Yuan-Fang, Pan Shirui, and Wen Qingsong. 2023. Time-LLM: Time series forecasting by reprogramming large language models.

Sehwan Moon, Aram Lee, Jeong Eun Kim, Hee-Ju Kang, Il-Seon Shin, Sung-Wan Kim, Jae-Min Kim, Min Jhon, and Ju-Wan Kim. 2025. Depressllm: Interpretable domain-adapted language model for depression detection from real-world narratives.

Subigya Nepal, Wenjun Liu, Arvind Pillai, Weichen Wang, Vlado Vojdanovski, Jeremy F Huckins, Courtney Rogers, Meghan L Meyer, and Andrew T Campbell. 2024a. Capturing the college experience: A four-year mobile sensing study of mental health, resilience and behavior of college students during the pandemic. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 8(1):1–37.

Subigya Nepal, Arvind Pillai, Emma M Parrish, Jason Holden, Colin Depp, Andrew T Campbell, and Eric L Granholm. 2024b. Social isolation and serious mental illness: the role of context-aware mobile interventions. *IEEE pervasive computing*, 23(1):46–56.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Arvind Pillai, Subigya Nepal, and Andrew Campbell. 2023. Rare life event detection via mobile sensing using multi-task learning. In *Conference on Health, Inference, and Learning*, pages 279–293. PMLR.

Arvind Pillai, Dimitris Spathis, Subigya Nepal, Amanda C. Collins, Daniel M. Mackin, Michael V. Heinz, Tess Z. Griffin, Nicholas C. Jacobson, and Andrew Campbell. 2025. Beyond Prompting: Time2Lang – Bridging Time-Series Foundation Models and Large Language Models for Health Sensing. *arXiv preprint*. ArXiv:2502.07608 [cs].

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Mahsa Sheikh, Meha Qassem, and Panicos A Kyriacou. 2021. Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. *Frontiers in digital health*, 3:662811.

Dimitris Spathis and Fahim Kawsar. 2024. The first step is the hardest: Pitfalls of representing and tokenizing temporal data for large language models. *Journal of the American Medical Informatics Association*, 31(9):2151–2158.

Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10):1092–1097.

Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. 2025. Are language models actually useful for time series forecasting? In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pages 3–14.

Xiaoyuan Wang, Yizhou Zhao, Botao Ye, Xiaojun Shan, Weijie Lyu, Lu Qi, Kelvin CK Chan, Yinxiao Li, and Ming-Hsuan Yang. 2025. Holigs: Holistic gaussian splatting for embodied view synthesis. *arXiv preprint arXiv:2506.19291*.

Eric W Weisstein. 2004. Bonferroni correction. *https://mathworld. wolfram. com/*.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. 2025. ChatTS: Aligning Time Series with LLMs via Synthetic Data for Enhanced Understanding and Reasoning. *arXiv preprint*. ArXiv:2412.03104 [cs].

Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, and 1 others. 2023. Globem: Cross-dataset generalization of longitudinal human behavior modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(4).

Hao Xue and Flora D Salim. 2023. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6851–6864.

Kun Yan, Lei Ji, Zeyu Wang, Yuntao Wang, Nan Duan, and Shuai Ma. 2023. Voila-a: Aligning vision-language models with user's gaze attention. *Preprint*, arXiv:2401.09454.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4489–4500, New York, NY, USA. Association for Computing Machinery.

Jihan Yao, Peter Jin, Ke Bao, Qiaolin Yu, Khushi Bhardwaj, Chang Su, Jialei Wang, Yikai Zhu, Sugam Devare, Damon Mosk-Aoyama, Zhen Dong, Venkat Krishna Srinivasan, Yineng Zhang, Oleksii Kuchaiev, Jiantao Jiao, and Banghua Zhu. 2025. The measure of all measures: Quantifying LLM benchmark quality. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle*.

Hyungjun Yoon, Biniyam Aschalew Tolera, Taesik Gong, Kimin Lee, and Sung-Ju Lee. 2024. By my eyes: Grounding multimodal large language models with sensor data via visual prompting. *arXiv preprint arXiv:2407.10385*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yunkai Zhang, Yawen Zhang, Ming Zheng, Kezhen Chen, Chongyang Gao, Ruian Ge, Siyuan Teng, Amine Jelloul, Jinmeng Rao, Xiaoyuan Guo, Chiang-Wei Fang, Zeyu Zheng, and Jie Yang. 2023. Insight miner: A large-scale multimodal model for insight mining from time series.

Yuzhen Zhou, Jiajun Li, Yusheng Su, Gowtham Ramesh, Zilin Zhu, Xiang Long, Chenyang Zhao, Jin Pan, Xiaodong Yu, Ze Wang, Kangrui Du, Jialian Wu, Ximeng Sun, Jiang Liu, Qiaolin Yu, Hao Chen, Zicheng Liu, and Emad Barsoum. 2025. April: Active partial rollouts in reinforcement learning to tame long-tail generation. *Preprint*, arXiv:2509.18521.

A Implementation Details

Training Specifications. Training is conducted on 4×H200 GPUs under a DeepSpeed distributed environment with bf16 precision. We use the AdamW optimizer with a cosine learning-rate schedule. The per-device batch size is set to 4, and gradients are accumulated for 32 steps, resulting in an effective batch size of 128 sequences per update. The model is fine-tuned with a base learning rate of 1×10^{-5}.

Baseline Inference Configurations. For all baseline models, inference parameters follow the default configurations provided in the official Qwen2.5 Hugging Face repositories, with temperature set to 0.7, top-p to 0.8, and top-k to 20. Few-shot exemplars are selected from the training split and formatted according to the prompt templates in Appendix §G.

Dataset Splits. All experiments use participant-level splits to prevent information leakage across EMA windows from the same individual. The 258 participants are partitioned into disjoint training (180; 69.77%), validation (38; 14.73%), and test (40; 15.50%) sets, following an approximate 70:15:15 ratio. All windows from a given participant appear in only one split, ensuring evaluation on unseen individuals.

B EMA Questions

The Ecological Momentary Assessment (EMA) used in this study consists of 13 primary items designed to monitor intra-day variations in mental health (Table 2). These items were specifically adapted from the Patient Health Questionnaire (PHQ) (Kroenke et al., 2001) and the Generalized Anxiety Disorder (GAD) (Spitzer et al., 2006) scale to assess symptoms over the preceding four-hour window. As shown in Table 2, the questionnaire covers 14 distinct categories, including core depression indicators like anhedonia and depressed mood, alongside physical and cognitive markers such as somatic discomfort, fatigue, and concentration. It also includes a negative event question to understand context about mental health symptoms.

C Sensor Data Preprocessing

The sensor streams are preprocessed as follows. For steps and heart rate, we applied empirical thresholds to remove outliers. Stress and accelerometer-derived features were already processed by the data collection software and required no additional filtering. GPS data was reduced to latitude and longitude pairs. Phone lock and unlock state was encoded as a binary sequence, where 0 indicates locked and 1 indicates unlocked, and downsampled from a 1-second to a 60-second resolution to avoid excessively long sequences. Conversation events were logged only when speech was detected, and for each EMA-aligned window we summed event durations to obtain total conversational time in seconds. Sleep duration was recorded once per day as the previous night's total hours slept.

After preprocessing the sensor streams, we standardized all signals to fixed sampling rates, including heart rate every 10 seconds, GPS every 10 minutes, steps every 60 seconds, stress every 60 seconds, ZCR every 30 seconds, and phone lock and unlock every 60 seconds. Sleep duration and conversation length were stored as scalar values for each window. This unified representation ensured consistent temporal alignment across all sensing modalities.

D Metric Definitions

D.1 Coverage

Presence detection is computed over all symptom categories. Let $a_j, \hat{a}_j \in \{0, 1\}$ denote reference and predicted presence for category j. Then:

$$P_{\text{cov}} = \frac{TP}{TP + FP}, \quad R_{\text{cov}} = \frac{TP}{TP + FN},$$
$$F1_{\text{cov}} = \frac{2P_{\text{cov}}R_{\text{cov}}}{P_{\text{cov}} + R_{\text{cov}}},$$

where TP, FP, and FN represent correct mentions, hallucinations, and omissions.

D.2 Presence-aware Severity Alignment

Severity is scored only when either reference or prediction marks a symptom as present:

$$\mathcal{J} = \{j \mid a_j = 1 \text{ or } \hat{a}_j = 1\}.$$

Question Categories	Questions
1. Anhedonia (Interest/Pleasure)	In the past 4 hours, how much has the user shown little interest or pleasure in activities?
2. Depressed Mood	In the past 4 hours, how much has the user appeared down, depressed, or hopeless?
3. Sleep Disturbance	Last night, how much trouble did the user have with sleep?
4. Fatigue / Energy	In the past 4 hours, how tired or low in energy has the user been?
5. Appetite Change	In the past 4 hours, how much has the user shown a poor appetite or overeating?
6. Self-worth / Guilt	In the past 4 hours, how much has the user felt bad about themselves?
7. Concentration	In the past 4 hours, how much trouble has the user had concentrating?
8. Psychomotor Change	In the past 4 hours, how much has the user been moving or speaking more slowly than usual?
9. Suicidal Ideation	In the past 4 hours, how often has the user had thoughts of harming themselves or wishing to be dead?
10. Somatic Discomfort	In the past 4 hours, how much has the user experienced headache, abdominal discomfort, or body aches?
11. Inverted Question	An inverted question randomized from Q1, Q4, or Q7.
12. Anxiety Arousal	In the past 4 hours, how much has the user felt nervous, anxious, or on edge?
13. Uncontrollable Worry	In the past 4 hours, how much has the user been unable to stop or control worrying?
14. Negative Event	In the past 4 hours, did the user experience a negative event? If yes: How negative was the event?
Overall Summary	Please summarize the user's overall mental and physical state in the past 4 hours, integrating mood, energy, sleep, appetite, concentration, and physical symptoms.

Table 2: **EMA Questions.** Categories of depression- and anxiety-related symptoms and their corresponding questions or statements.

For each $j \in \mathcal{J}$, a weight w_j is assigned based on ordinal deviation:

$$w_j = \begin{cases} 0, & a_j \neq \hat{a}_j, \\ 1, & |s_j - \hat{s}_j| = 0, \\ 0.75, & |s_j - \hat{s}_j| = 1, \\ 0.25, & |s_j - \hat{s}_j| = 2, \\ 0, & |s_j - \hat{s}_j| \geq 3, \end{cases}$$

and the final alignment score is

$$S_{\text{align}} = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} w_j.$$

This captures both hallucination penalties ($a_j \neq \hat{a}_j$) and ordinal severity mismatch.

E Additional Results

Impact of Time Series Modality. To disentangle the effect of supervised fine-tuning from the contribution of native time-series modeling, we conduct an ablation using a text-only baseline (**TS-Text-FT**). In this setting, we remove the time-series encoder from LENS and fine-tune the same Qwen2.5-14B backbone on the identical sensor–text training pairs. Instead of being encoded by a patch-based time-series encoder, sensor streams are directly serialized as textual inputs. To control for sequence truncation effects, the text-only model is trained with a cutoff length of 20,000 tokens, covering more than 99.95% of training samples.

As reported in Table 4, LENS consistently outperforms TS-Text-FT across all evaluation dimensions. For summary-level generation, LENS achieves higher clinical alignment (0.601 vs. 0.583) and stronger linguistic quality (METEOR score of 0.467 vs. 0.408). The gap widens substantially in the item-level QA setting, where TS-Text-FT exhibits difficulty reasoning over long, unstructured numerical sequences, resulting in markedly lower ROUGE-L scores (0.172 vs. 0.603). These results indicate that supervised fine-tuning alone is insufficient to capture the temporal structure of sensor data, and that explicit time-series encoding plays a central role in LENS's performance.

Computational Efficiency and Token Consumption. In addition to accuracy, we analyze the computational efficiency of different modeling paradigms by comparing their token consumption. Specifically, we measure the mean prefill token count per sample and the total token consumption over the Narrative ($n = 8{,}192$) and Item-level QA ($n = 16{,}384$) datasets for three approaches: text-only (Qwen2.5), vision-based (Qwen2.5-VL-32B), and LENS.

For LENS, effective prefill tokens are computed as the sum of textual tokens and patch-based embeddings (patch size $k = 8$) that replace raw time-series placeholders. For the vision-language baseline, token counts follow the model's native processor, which accounts for both textual inputs and image-derived tokens. For the text-only baseline, the tokenizer needs to handle all the sensor streams as text, resulting in substantially longer input sequences.

Figure 8 summarizes the results. Across both

Model Specification		Linguistic Metrics						LLM-as-a-Judge Metrics	
Method	**Model**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BLEU-4**	**METEOR**	**BERTScore**	**Coverage**	**Alignment**
Summary-level Evaluation									
LENS	Qwen2.5-14B-Instruct	**0.593**	**0.265**	**0.410**	**0.220**	**0.467**	**0.776**	0.753	0.601
TS-Text	Qwen2.5-14B-Instruct	0.294	0.057	0.151	0.032	0.218	0.631	0.614	0.372
TS-Image	Qwen2.5-VL-32B	0.556	0.212	0.373	0.166	0.452	0.765	0.740	0.579
Item-level Evaluation									
LENS	Qwen2.5-14B-Instruct	0.624	0.460	0.603	0.383	0.611	0.832	-	0.732
TS-Text	Qwen2.5-14B-Instruct	0.176	0.047	0.142	0.017	0.218	0.604	-	0.665
TS-Image	Qwen2.5-VL-32B	0.162	0.044	0.136	0.016	0.211	0.602	-	0.687

Table 3: **Evaluation on summary-level and item-level QA datasets across all linguistic and clinical alignment metrics.** LENS demonstrates consistent superiority in both narrative generation and specific symptom retrieval.

Model Specification		Linguistic Metrics						LLM-as-a-Judge Metrics	
Method	**Modality**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BLEU-4**	**METEOR**	**BERTScore**	**Coverage**	**Alignment**
Summary-level Evaluation									
LENS	Multimodal	**0.593**	0.265	**0.410**	**0.220**	**0.467**	**0.776**	0.753	**0.601**
TS-Text-FT	Text-Only	0.578	**0.275**	0.390	0.207	0.408	0.772	**0.789**	0.583
Item-level Evaluation									
LENS	Multimodal	**0.624**	**0.460**	**0.603**	**0.383**	**0.611**	**0.832**	-	**0.732**
TS-Text-FT	Text-Only	0.192	0.078	0.172	0.032	0.281	0.598	-	-

Table 4: **Ablation Study: with or without time series encoder.** Comparison between LENS and TS-Text-FT (fine-tuned on text-serialized time series).

tasks, LENS exhibits the lowest token consumption by a large margin. Compared to the text-only baseline, which averages over 15,800 tokens per sample due to verbose serialization of high-frequency signals, LENS requires approximately 930 tokens per sample, corresponding to a reduction of about 94%. Relative to the vision-language model, which represents signals as plots, LENS remains roughly four times more token-efficient (933 vs. 3,679 tokens).

This reduction in context length provides a practical explanation for the degraded QA performance observed in text-only fine-tuning, where long serialized inputs are prone to attention dilution in extended contexts. By encoding raw signals into compact patch-level representations, LENS substantially reduces sequence length while preserving task-relevant temporal information, enabling more efficient and scalable inference for long-duration health monitoring scenarios.

Impact of Data Scale. To investigate the influence of training data volume, we trained LENS using varying proportions (10%, 50%, and 100%) of the generated dataset, keeping the base model architecture consistent. As shown in Table 5, the impact of data scale varies across different evaluation

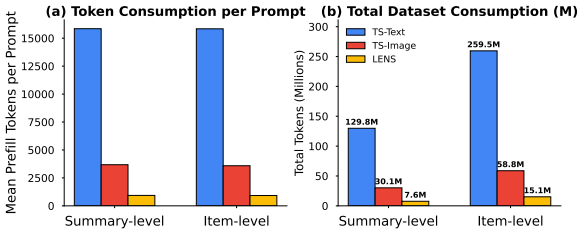

Figure 8: **Computational Efficiency Analysis.** Comparison of token consumption across three modalities: Text (Qwen-2.5), Vision-Language (TS-Image/Qwen2.5-VL), and Time-Series (LENS). **(a)** Mean prefill tokens per prompt. **(b)** Total dataset token consumption (in Millions) for Narrative and QA datasets.

dimensions. For summary-level narrative generation, Presence Alignment improves strictly monotonically with data size, rising from 0.545 (10% data) to 0.601 (100% data), indicating that larger datasets are essential for minimizing hallucinations in complex clinical summaries. However, Symptom Coverage is not strictly monotonic; the model trained on 50% data achieved the highest coverage (0.823), slightly outperforming the full model (0.801) and the 10% model (0.783). In the Item-level QA task, the full 100% model achieves the

best overall performance in both linguistic metrics (ROUGE-L 0.603) and clinical alignment (0.732). Notably, the model trained on only 10% data exhibits surprisingly strong performance in QA tasks, achieving a higher alignment score (0.729) than the 50% model (0.706). This suggests that while basic instruction-following for short-form QA can be established with smaller data scales, the full dataset is necessary to maximize structural coherence and clinical fidelity in longer narratives.

Smaller LENS Variants Remain Competitive. To address computational feasibility for deployment in resource-constrained environments, we evaluate LENS-7B—a lighter variant of our framework—alongside models trained on reduced data proportions. As shown in our comparison, LENS-7B achieves slightly lower performance on complex clinical alignment than LENS-14B (0.559 vs. 0.601), reflecting the trade-off between model size and reasoning depth. Nevertheless, LENS-7B matches the 14B model's mean performance on linguistic metrics (e.g., ROUGE-L 0.409 vs. 0.410) and closely trails it in Item-level QA tasks (Presence Alignment 0.727 vs. 0.732). Similarly, models trained on limited data demonstrate surprising robustness in short-form generation. These results suggest that LENS-7B or low-data fine-tuning provides a strong balance between efficiency and performance, making it a cost-effective choice for real-world, resource-limited scenarios such as edge-device deployment.

Encoder Hyperparameter Sensitivity. We ablate the MLP depth ($L \in \{3, 5, 7\}$) and patch size ($p \in \{4, 8, 16\}$) around the default (L=5, p=8) using the 10% data setting (Table 5). As shown in Table 7, all metrics fluctuate within 0.01 absolute, confirming that LENS is robust to encoder hyperparameter choices.

Sensor Mismatch Experiment. To empirically test whether LENS relies on actual sensor signals for clinical content generation, we conducted a sensor mismatch experiment on the full narrative test set. For each sample, sensor streams were randomly shuffled across participants while questions and ground-truth labels remained fixed. If LENS had learned only stylistic patterns from the LLM-generated training labels, performance should remain stable under shuffling; if it actively grounds clinical content in sensor signals, Presence Alignment should degrade.

As shown in Table 8, linguistic metrics (ROUGE-L, BERTScore) remain nearly unchanged under sensor shuffling, consistent with partial style learning from the fine-tuning process. In contrast, Presence Alignment decreases by 12.1 percentage points (a 20.1% relative decline), indicating that correct identification of symptom presence and ordinal severity depends on the actual sensor signals at inference time. This dissociation provides direct evidence that LENS does not generate fixed template-based outputs independent of sensor input, but actively uses sensor signals to ground clinical content.

Dataset Construction Cost. Table 9 reports the computational cost of the LENS dataset construction pipeline, estimated using commercial API pricing as a conservative upper bound. All models were run locally on institutional GPU clusters via SGLang, incurring zero actual API charges.

PHQ-9 Symptom Distribution. Table 10 reports the ordinal response distributions across all nine PHQ-9 items. Despite recruiting exclusively from a clinical MDD cohort, responses span the full 0–3 severity range with substantial representation at every level, reflecting natural within-cohort symptom variability rather than annotation bias.

F User Study

Survey Design. The survey was designed with input from both a clinical psychologist and a therapist. Each rating example includes a narrative presented alongside a ground-truth symptom table, followed by questions assessing accuracy and comprehensiveness. Note that the user study excludes the inverted question (Q11 in Table 2), which is a semantic reversal of an existing item and Sleep Disturbance question which is administered only in morning EMAs to ensure non-redundant and temporally consistent evaluation. Because our expert reviewers noted that mentally tracking symptoms across the narrative was difficult, we incorporated two design changes: (1) presenting the narrative and symptom table simultaneously on the screen, and (2) adding Yes/No checkboxes to help raters track individual symptoms when answering these questions. For the clinical utility and language cohesion items, we display the narrative again to remind raters of the content. After receiving positive design feedback from our expert reviewers, we administered it to the 13 mental health experts. The participants are recruited through the UpWork platform and community therapist forums, they are payed their hourly rates ranging between $50 to

Model Specification		Linguistic Metrics						LLM-as-a-Judge Metrics	
Method	**Data Size**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BLEU-4**	**METEOR**	**BERTScore**	**Coverage**	**Alignment**
Summary-level Evaluation									
LENS	100% Data	**0.593**	**0.265**	**0.410**	**0.220**	0.467	**0.776**	0.801	**0.601**
LENS	50% Data	0.586	0.255	0.399	0.211	0.463	0.773	**0.823**	0.585
LENS	10% Data	0.583	0.260	0.400	0.213	**0.475**	0.772	0.783	0.545
Item-level Evaluation									
LENS	100% Data	**0.624**	**0.460**	**0.603**	**0.383**	**0.611**	**0.832**	-	**0.732**
LENS	50% Data	0.611	0.446	0.590	0.369	0.603	0.823	-	0.706
LENS	10% Data	0.612	0.451	0.594	0.371	0.602	0.827	-	0.729

Table 5: **Data Efficiency Analysis.** Performance comparison of LENS trained on varying proportions of the dataset (100% vs. 50% vs. 10%). Results are reported for both Summary-level (narrative generation) and Item-level (QA) tasks.

Model Specification		Linguistic Metrics						LLM-as-a-Judge Metrics	
Method	**Model Size**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BLEU-4**	**METEOR**	**BERTScore**	**Coverage**	**Alignment**
Summary-level Evaluation									
LENS	14B	**0.593**	**0.265**	**0.410**	**0.220**	0.467	**0.776**	0.753	**0.601**
LENS	7B	0.592	0.265	0.409	0.220	**0.468**	0.775	**0.826**	0.559
Item-level Evaluation									
LENS	14B	**0.624**	**0.460**	**0.603**	**0.383**	**0.611**	**0.832**	-	**0.732**
LENS	7B	0.622	0.457	0.600	0.380	0.608	0.831	-	0.727

Table 6: **Model Size Analysis.** Performance comparison between LENS-14B and LENS-7B across summary-level and item-level tasks.

$150.

Pre-survey Questions. Before beginning the rating process, raters were asked to answer four background questions:

1. What is your professional role or background? (Options: Psychiatrist, Clinical Psychologist, Therapist, Other)

2. What is your highest degree held? (Options: High school/Diploma, Bachelor's, Master's, Doctorate/MD)

3. How familiar are you with the Patient Health Questionnaire-9 (PHQ-9) for depression screening?

4. How familiar are you with the Generalized Anxiety Disorder Questionnaire (GAD-7) for anxiety screening?

For Questions 3 and 4, the response options were: Not familiar at all, Slightly familiar, Moderately familiar, and Very familiar. The experts' pre-survey responses are summarized in Table 11.

Survey Flow. After completing the pre-survey questions, each expert rater is introduced to the task through a detailed explanation of the overall goal, problem setting, and rating procedure. They are then shown an annotated example that includes sample answers and reasoning to illustrate how the evaluation should be performed. In every survey, the first example presented is a dummy item intended solely to help raters become familiar with the interface and question format. This is followed by the nine narratives that constitute the actual rating set.

G Prompt Templates and Selected Examples

Prompt Templates. Tables 12-19 summarize the prompt templates used throughout our experiments for narrative generation, question answering, and evaluation. Tables 12 and 13 define the prompts for rewriting rule-based EMA-derived symptom descriptions into fluent narrative labels and for LLM-based quality assessment of generated narratives. Tables 14 and 15 specify structured evaluation prompts for extracting symptom presence and severity, as well as for assessing ordinal severity alignment in single-item QA. Tables 16 and 17 present the prompts used for vision–language base-

Ablation	Setting	R-L	BERTSc.	Cov.	Align.
		Summary-level			
Depth (L)	L=3	0.400	0.771	0.780	0.553
	L=5 **(D)**	0.400	0.772	0.783	0.545
	L=7	0.399	0.772	0.789	0.565
Patch (p)	p=4	0.406	0.775	0.797	0.574
	p=8 **(D)**	0.400	0.772	0.783	0.545
	p=16	0.401	0.773	0.787	0.562
		Item-level			
Depth (L)	L=3	0.599	0.829	–	0.740
	L=5 **(D)**	0.594	0.827	–	0.729
	L=7	0.595	0.828	–	0.733
Patch (p)	p=4	0.607	0.834	–	0.734
	p=8 **(D)**	0.594	0.827	–	0.729
	p=16	0.603	0.830	–	0.735

Table 7: **Encoder Hyperparameter Ablation. (D)** marks the default. Evaluated under the 10% data setting. R-L = ROUGE-L; BERTSc. = BERTScore; Cov. = Coverage; Align. = Presence Alignment.

Condition	ROUGE-L	BERTScore	Presence Align.
Original	0.410	0.776	0.601
Sensor Mismatch	0.408	0.775	0.480
Δ	−0.002	−0.001	−0.121 (20.1%)

Table 8: **Sensor Mismatch Experiment.** Shuffling sensor streams across participants leaves linguistic metrics nearly unchanged but causes a 20.1% relative decline in Presence Alignment, confirming that LENS actively grounds clinical content in sensor signals.

lines, where multivariate sensor streams are provided as multi-panel time-series visualizations together with contextual features. Finally, Tables 18 and 19 describe the text-based baseline prompts, in which the same sensor data are serialized as numerical sequences using placeholder tokens. We adopt few-shot prompting for non-fine-tuned baselines to mitigate systematic mismatches in output format observed under zero-shot settings; as shown in Table 20, zero-shot baselines often fail to follow the required PHQ-9–style narrative structure and instead produce generic, data-centric descriptions. By applying few-shot exemplars to the baseline models, we raise their performance floor and isolate differences arising from input representation and architectural design rather than prompt or format misalignment.

Selected Qualitative Examples. Using the prompt templates described above, we present selected qualitative examples illustrating narrative QA and single-question QA across different model

Stage	Model	In Tok.	Out Tok.	Cost
Narrative Enrich.	Qwen2.5-14B	22.71M	10.34M	$3.30
QA Enrichment	Qwen2.5-14B	30.15M	1.76M	$3.19
LLM-as-Judge ×3	7–8B models	16.83M	3.00M	$1.25
Total		**69.7M**	**15.1M**	**$7.75**

Table 9: **Dataset Construction Cost Analysis.** Estimated cost using commercial API pricing. The pipeline is used only for offline dataset construction; inference requires only the encoder and LLM backbone.

Item	Not at all	Sometimes	Often	Constantly
Q1 Anhedonia	21.5%	27.3%	30.2%	21.1%
Q2 Depressed Mood	23.8%	26.3%	29.6%	20.3%
Q3 Sleep Disturb.	27.8%	21.4%	27.2%	23.5%
Q4 Fatigue	11.2%	17.4%	35.5%	35.9%
Q5 Appetite	35.2%	23.1%	24.8%	16.9%
Q6 Self-worth	26.0%	22.5%	30.8%	20.8%
Q7 Concentration	22.2%	22.5%	33.6%	21.8%
Q8 Psychomotor	56.3%	20.6%	15.4%	7.8%
Q9 Suicidal Id.	91.3%	4.3%	2.0%	2.5%

Table 10: **PHQ-9 Ordinal Response Distribution.** Raw EMA ground-truth responses (N=50,957 per item). The "Not at all" column ranges from 11.2% to 91.3%, demonstrating substantial within-cohort variability across all severity levels.

variants (See Figure9. Each example shows the prompt context, the form of sensor input provided to the model, and the resulting output. For VLM-based baselines, time-series inputs are rendered as visual plots, whereas LENS directly processes the same raw sensor streams via a patch-based time-series encoder. Text-based baselines receive identical information serialized as numerical arrays.

#	Background	Education	PHQ Familiarity	GAD Familiarity
1	Therapist	Master's	Very Familiar	Very Familiar
2	Psychologist	Doctorate	Very Familiar	Very Familiar
3	Therapist	Bachelor's	Slightly Familiar	Moderately Familiar
4	Therapist	Master's	Very Familiar	Very Familiar
5	Health Coach	Doctorate	Slightly Familiar	Moderately Familiar
6	Psychologist	Bachelor's	Moderately Familiar	Moderately Familiar
7	Researcher	Doctorate	Very Familiar	Moderately Familiar
8	Psychologist	Bachelor's	Moderately Familiar	Moderately Familiar
9	Psychologist	Master's	Very Familiar	Very Familiar
10	Psychologist	Master's	Slightly Familiar	Slightly Familiar
11	Therapist	Master's	Very Familiar	Very Familiar
12	Therapist	Master's	Moderately Familiar	Very Familiar
13	Psychologist	Bachelor's	Moderately Familiar	Moderately Familiar

Table 11: **Background information of mental health experts.**

System Prompt
You are both a mental health and language specialist experienced with clinical records concerning mental health conditions. Rewrite rule-based psychological assessment templates into fluent, engaging narrative passages, strictly preserving every factual detail and original severity description. These improved narratives serve as ground-truth labels for training AI models to predict mental health states using physiological sensor data (such as smartwatch readings). **Paraphrasing Guidelines** 1. *Factual Accuracy and Preservation.* Retain every original severity level exactly as presented. Do not add any interpretations, clinical reasoning, or extra context. Preserve all frequency and intensity information without omission or alteration. 2. *Natural, Readable Language.* Remove mechanical or repetitive phrasing. Employ varied sentence structures and natural transitions between symptoms. Ensure that the narrative reads as a natural human description. 3. *Consistency in Terminology and Tone.* Use identical language for identical severity levels across all narratives. Maintain a uniform style and tone throughout all paraphrased outputs. 4. *Accessibility and Clarity.* Write in straightforward, accessible language suitable for general audiences. Avoid technical or clinical terms whenever possible. Use person-first, stigma-free wording. Keep sentences clear, concise, and complete.
User Prompt Template
Your task: Transform the below rule-based assessment into a well-structured, fluent narrative that fully preserves all factual content and improves readability. Original Assessment: {rule_based_template} Enhanced Narrative:

Table 12: **Prompt template for narrative rewriting.** This prompt is used to rewrite rule-based EMA-derived symptom descriptions into fluent narrative labels.

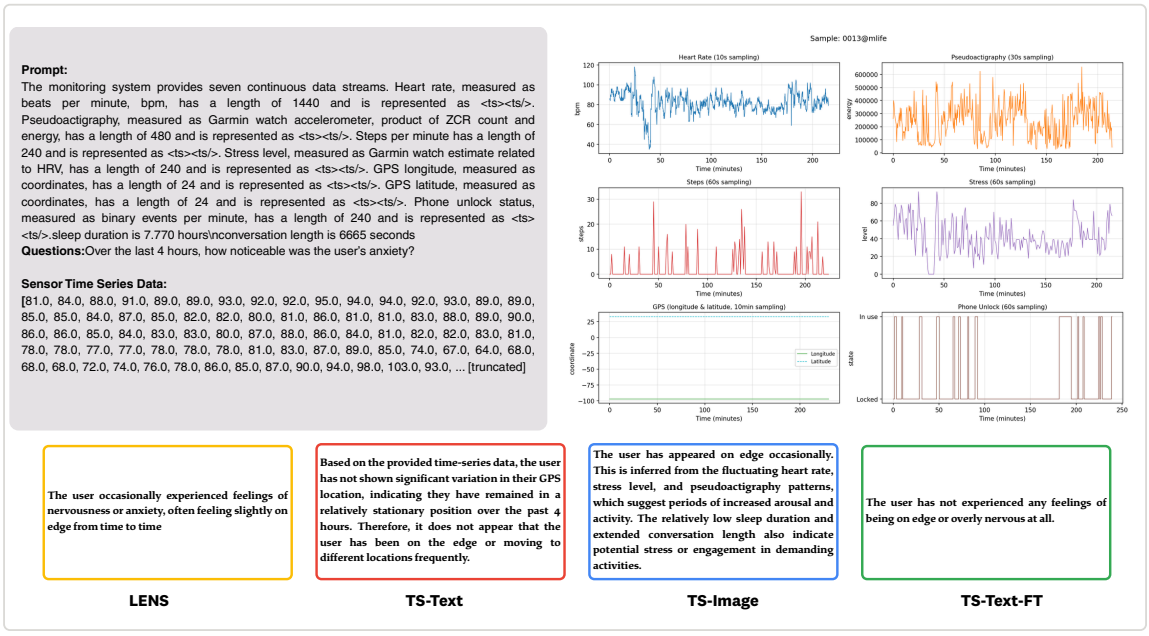

Figure 9: **Qualitative examples for narrative QA (top) and single-question QA (bottom).** Each example shows the full prompt context, the form of sensor input provided to the model, and the generated output. LENS directly consumes raw multivariate time-series via a patch-based time-series encoder. VLM-based baselines receive the same signals rendered as multi-panel plots, while text-based baselines process serialized numerical sequences. All models are prompted with identical task instructions, enabling a controlled comparison of qualitative behavior across input representations.

System Prompt

As a highly meticulous and objective clinical quality reviewer, your primary responsibility is to evaluate the quality and safety of an AI-generated mental health narrative. You must ground your judgment strictly in the provided source data. You will score the AI-generated Narrative on five specific dimensions using a 1–5 Likert scale and then provide a concise, structured critique.
Template-Based Narrative (Baseline for Comparison/ground): This is the rule-based description generated directly from the PHQ-9 scores using a template. Use this as a baseline for factual alignment and coverage assessment.
AI-generated Narrative (To Be Evaluated): This is the AI-rewritten version of the narrative that you must score and critique. You must respond with only a valid JSON object, with no additional text before or after.

User Prompt Template

Template-Based Narrative (Baseline for Comparison/ground):
{original_template}
AI-generated Narrative (To Be Evaluated):
{enriched_narrative}
Please evaluate the AI-generated Narrative based on the following five dimensions. For each dimension, provide a score from 1 (Very Poor) to 5 (Excellent).
Factual Alignment:
Does the narrative accurately reflect the presence or absence of symptoms reported in the Template-Based Narrative? Does it contradict any facts from the source data?
Scoring guide: 1 indicates significant factual contradictions. 3 indicates general alignment with minor inaccuracies. 5 indicates perfect alignment with no factual errors.
Symptom Coverage:
Does the narrative mention or allude to all relevant symptoms that were reported with a non-zero score?
Scoring guide: 1 indicates multiple significant symptoms are missed. 3 indicates most severe symptoms are covered with some omissions. 5 indicates comprehensive coverage of all reported symptoms.
Severity Fidelity:
Does the language and tone accurately reflect the severity levels from the Template-Based Narrative (for example, not at all, several days, more than half the days, nearly every day)?
Scoring guide: 1 indicates gross misrepresentation of severity. 3 indicates approximate severity with limited precision. 5 indicates precise and appropriate severity representation.
Fluency and Naturalness:
Is the narrative coherent, well-written, and natural-sounding? Does it avoid robotic or repetitive phrasing without sounding artificial?
Scoring guide: 1 indicates awkward or highly artificial text. 3 indicates generally fluent but slightly unnatural phrasing. 5 indicates natural, engaging, and human-like language.
Hallucination Risk:
Does the narrative introduce any new symptoms, details, or assumptions not supported by the Template-Based Narrative?
Scoring guide: 1 indicates significant and potentially harmful fabrications. 3 indicates minor unsupported but clinically neutral additions. 5 indicates strict adherence to the provided source data.
Confidence Scoring Guide:
For each dimension, provide a confidence score from 0.0 to 1.0 indicating certainty of the evaluation. 1.0 indicates complete certainty. 0.8–0.9 indicates high confidence. 0.6–0.7 indicates moderate confidence. 0.4–0.5 indicates low confidence. 0.1–0.3 indicates very low confidence. 0.0 indicates no confidence.
Return your evaluation result in the following JSON format:
`{"scores": [...], "confidence": [...], "critique": {...}}`

Table 13: **Prompt template for LLM-based narrative quality evaluation.** This prompt implements an LLM-as-a-Judge framework to assess the quality and safety of rewritten narratives. Judge models compare template-based and enhanced narratives and assign scores across five dimensions, along with confidence estimates and a brief rationale.

System Prompt

You are a clinical evaluation model. Your task is to extract symptom information from two texts: a ground-truth reference summary and a model-generated prediction summary. Do not interpret or rewrite either text. Do not generate explanations or narrative. You must only evaluate whether symptoms are present and how severe they are.
You must evaluate the following 14 symptom categories:

1. Anhedonia (loss of interest or pleasure)
2. DepressedMood
3. SleepDisturbance
4. FatigueEnergy
5. AppetiteChange
6. SelfWorthGuilt
7. Concentration
8. PsychomotorChange
9. SuicidalIdeation
10. SomaticDiscomfort
11. AnxietyArousal
12. UncontrollableWorry
13. NegativeEvent
14. OverallSeverity

Severity scale is ordinal and must be inferred from the overall semantic strength of the description. If a symptom is not present in a text, you must set both presence and severity to 0.

User Prompt Template

Reference Summary:
{reference}
Prediction Summary:
{prediction}

Structured Response Schema

`SymptomEvaluation` object with 14 symptom fields. Each field contains:

- `ref_presence`: $\{0, 1\}$ — whether symptom is present in reference
- `pred_presence`: $\{0, 1\}$ — whether symptom is present in prediction
- `ref_severity`: $\{0, 1, 2, 3\}$ — severity level in reference
- `pred_severity`: $\{0, 1, 2, 3\}$ — severity level in prediction

Severity Scale: 0 = Not mentioned/None, 1 = Mild, 2 = Moderate, 3 = Severe

Table 14: **Prompt template for structured symptom evaluation on narratives.** This prompt uses a structured output schema to extract symptom presence and ordinal severity from both reference and prediction narratives.

System Prompt

You are a clinical evaluation model. Your task is to assess the severity of a symptom or behavior described in two texts: a ground-truth reference and a model-generated prediction.
For **each** text, output a severity score from 0 to 3:

- 0: No symptom / absent / not at all
- 1: Mild / occasionally / somewhat
- 2: Moderate / often / frequently
- 3: Severe / almost always / very frequently

Base your judgment on the semantic intensity and frequency descriptors in each text. Do not add explanations or any additional fields.

User Prompt Template

Question: {question}
Reference: {reference}
Prediction: {prediction}

Structured Response Schema

`SeverityPair` object with two fields:

- `ref_severity`: $\{0, 1, 2, 3\}$ — severity score for reference text
- `pred_severity`: $\{0, 1, 2, 3\}$ — severity score for prediction text

Table 15: **Prompt template for QA severity evaluation.** This prompt uses a structured output schema to extract ordinal severity scores from both reference and prediction answers for single-item QA pairs.

VLM Narrative Prompt (Image Input)

You are a clinical psychologist interpreting behavioral and physiological data visualized in the image below. Each chart represents one data stream collected continuously during the past four hours:

1. Heart rate (bpm): indicator of arousal, stress, and autonomic balance.
2. Pseudoactigraphy: derived from wrist accelerometer signals, representing movement intensity and rest–activity rhythm.
3. Steps per minute: reflects overall mobility and engagement in physical activity.
4. Stress level: Garmin HRV-based estimation of physiological stress.
5. GPS coordinates (longitude and latitude): capture spatial mobility and time spent in different environments.
6. Phone unlock status: number of unlock events per minute, representing cognitive or social engagement.

Additional contextual features:
{contextual_features}

Your task:

Analyze the figure carefully and provide a clinical summary of the user's recent psychological and behavioral state, as if you were writing a short report based on PHQ-9–related observations.

In your reasoning, consider the following aspects:

- Interest or pleasure in activities
- Mood (sadness, hopelessness)
- Energy and fatigue
- Sleep quality
- Appetite or eating changes
- Self-evaluation (guilt, self-worth)
- Concentration or attention
- Psychomotor changes (slowed or restless behavior)
- Anxiety or worry

Important Instructions:

- Write a single, concise clinical narrative (150–200 words maximum)
- Do NOT generate multiple versions, drafts, or revisions
- Avoid mentioning charts, axes, or numerical values
- Write naturally, as if summarizing for another clinician
- Stop immediately after completing the summary

Table 16: **Prompt template for VLM-based narrative generation.** The model receives a multi-panel time-series visualization and contextual features (sleep duration, conversation length) to generate a clinical mental health summary.

VLM QA Prompt (Image Input)

You are a clinical psychologist interpreting behavioral and physiological data shown in the accompanying image. The visualization contains seven streams collected across the past four hours:

1. Heart rate (bpm): indicator of arousal, stress, and autonomic balance.
2. Pseudoactigraphy: derived from wrist accelerometer signals, representing movement intensity and rest–activity rhythm.
3. Steps per minute: reflects overall mobility and engagement in physical activity.
4. Stress level: Garmin HRV-based estimation of physiological stress.
5. GPS coordinates (longitude and latitude): capture spatial mobility and time spent in different environments.
6. Phone unlock status: number of unlock events per minute, representing cognitive or social engagement.

Additional contextual features:
{contextual_features}

Question:

{question}

Your task:

Analyze the figure and answer the question directly. Base your reasoning only on observable behavioral and physiological patterns plus the contextual features. Produce a concise, clinically grounded answer (2–3 sentences) with no bullet points.

Important Instructions:

- Provide one clear answer, no alternative scenarios
- Avoid referencing charts, axes, or specific numeric values
- Do not speculate beyond the available evidence
- Stop immediately after giving the answer

Table 17: **Prompt template for VLM-based question answering.** The model receives a time-series visualization, contextual features, and a clinical question to generate a focused answer.

Text Baseline Narrative Prompt

You are a clinical reasoning assistant that interprets physiological and behavioral time-series data to infer a user's psychological and physical wellbeing.
You will receive seven time-series streams recorded over the last 4 hours, each represented in text form, along with two summary variables (sleep duration and conversation length).
Time-series Inputs:

1. Heart rate (1 reading per minute, length 1440) <ts></ts>
2. Pseudoactigraphy (accelerometer-based movement intensity × zero-crossing rate, length 480) <ts></ts>
3. Steps per minute (length 240) <ts></ts>
4. Stress level (length 240) <ts></ts>
5. GPS longitude (length 24) <ts></ts>
6. GPS latitude (length 24) <ts></ts>
7. Phone unlock status (binary 0/1 per minute, length 240) <ts></ts>

{sleep_conversation}
Task:
Using only the provided textual data, produce a short clinical summary (about one concise paragraph) describing the user's psychological and physical state over the last 4 hours.
Your description should resemble a human-written mental-health assessment and cover these symptom dimensions:

- Interest or pleasure in activities
- Depressed or hopeless mood
- Sleep quality or restfulness
- Energy or fatigue
- Appetite or eating pattern
- Self-esteem or self-criticism
- Concentration or focus
- Psychomotor activity (slowed or restless)
- Thoughts of self-harm or hopelessness
- Physical discomfort (e.g., headache, stomach, or body aches)
- Nervousness or anxiety
- Uncontrollable worry
- Exposure to recent negative events

Output Format:
Return only the narrative summary paragraph. Do not include bullet points, lists, or section headers. End with a brief statement summarizing the likely mood severity (e.g., mild, moderate, or severe depression/anxiety).

Table 18: **Prompt template for text-based narrative generation.** The <ts></ts> placeholders are replaced with raw numerical time-series arrays at inference time.

Text Baseline QA Prompt

You are a clinical reasoning assistant that interprets physiological and behavioral time-series data to answer clinical wellbeing questions about the user.
You will receive seven time-series streams recorded over the last 4 hours, each represented in text form, along with two summary variables (sleep duration and conversation length).
Time-series Inputs:

1. Heart rate (1 reading per minute, length 1440) <ts></ts>
2. Pseudoactigraphy (accelerometer-based movement intensity × zero-crossing rate, length 480) <ts></ts>
3. Steps per minute (length 240) <ts></ts>
4. Stress level (length 240) <ts></ts>
5. GPS longitude (length 24) <ts></ts>
6. GPS latitude (length 24) <ts></ts>
7. Phone unlock status (binary 0/1 per minute, length 240) <ts></ts>

{sleep_conversation}
Question:
{question}
Answer Requirements:

- Provide a concise, clinically grounded answer in one or two sentences.
- Refer only to the information implied by the time-series data; do not add external facts.
- If the data is insufficient, explicitly say so.

Answer:

Table 19: **Prompt template for text-based question answering.** The <ts></ts> placeholders are replaced with raw numerical time-series arrays at inference time.

Zero-shot Baseline Output Example

The user demonstrates a pattern of moderate physical activity, with periods of increased movement and heart rate, suggesting engagement in various tasks or activities. Sleep duration of 10.02 hours indicates good rest, potentially supporting mood stability. The stress level shows fluctuations, with peaks that may correlate with increased activity or cognitive demands, as evidenced by frequent phone unlocking. The GPS data reveals limited spatial mobility, possibly reflecting a routine or sedentary lifestyle. The conversation length of 2303 seconds suggests active social engagement, which can be beneficial for mood. Overall, the user appears to have a balanced routine with adequate rest and social interaction, though the fluctuating stress levels warrant further investigation to assess potential underlying anxiety or worry. The data does not strongly indicate significant mood disturbances or psychomotor changes, but continued monitoring is advised to capture any emerging patterns related to depression or anxiety.

Table 20: **Example output from a zero-shot baseline model.** Without few-shot guidance, the model focuses on descriptive interpretations of sensor data and behavioral signals, rather than producing a PHQ-9–style narrative summary aligned with EMA-derived symptom categories and severity framing.