Behavior Prediction from Everyday Sounds via LLMs with Multi-sensor Context and Priors

Anonymous Author(s)

Abstract

Audio serves as a high-density and non-intrusive modality for behaviour understanding but often lacks contextual grounding in complex environments. This paper proposes a modular framework that integrates raw audio, multi-sensor contextual signals (e.g., GPS, IMU, heart rate), and user priors, leveraging a multimodal large language model (MLLM) as the core for semantic and logical inference to achieve fine-grained user behaviour prediction. Based on preliminary user profiling, the system leverages data from everyday wearable devices—such as smartwatches—to generate fine-grained, structured records of daily behaviours, with the potential to support personalised feedback and habit tracking.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**.

Keywords

Behaviour recognition, large language models, multimodal sensing, audio understanding, user context modelling, wearable computing

ACM Reference Format:

Anonymous Author(s). 2018. Behavior Prediction from Everyday Sounds via LLMs with Multi-sensor Context and Priors. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Sound is an omnipresent and information-rich sensory modality that plays a crucial role in understanding human behaviour. From the sound of wind during a jog to the clatter of utensils during meals, humans can readily infer ongoing activities from auditory cues. Existing work demonstrates the strong potential of acoustic signals in behaviour understanding. Early methods used hand-crafted features like MFCCs and spectrograms with classical classifiers such as SVMs and HMMs [5]. With deep learning, these features were combined with CNNs and LSTMs to improve emotion and activity recognition [2, 4, 9]. More recently, large language models (LLMs) have shifted the focus from simple event classification to *semantic behaviour inference*. Multimodal LLMs like Gemini 1.5 Pro and SALMONN demonstrate unified reasoning over non-verbal audio, music, and emotion [6, 7], directly processing raw waveforms and supporting open-domain inference.

Yet, audio-only models often lack contextual grounding in real-world scenarios, leading to ambiguity. To address this, auxiliary data—including multimodal sensors (e.g., IMUs, GPS) and user-specific priors—has proven essential. Studies show that integrating these sources significantly boosts robustness and fine-grained inference for activities like eating and drinking [8, 11]. Surveys further highlight the role of demographic and behavioural priors in enabling personalised reasoning [3].

Motivated by this, we propose a unified behavior modeling framework that fuses audio, multimodal sensor data, and structured user knowledge, with LLMs serving as the central semantic reasoning engine. Our system captures audio from smartwatches as the primary modality, complemented by motion patterns (via IMU), geolocation (via GPS), and user-specific priors including demographics and behavior history. Existing studies have shown that LLMs face challenges when directly processing raw multimodal data, including limited accuracy, poor robustness, and weak temporal reasoning [1, 10, 12]. To address these limitations, we adopt a two-stage approach: transforming sensor signals into structured textual descriptions before passing them to the LLM. This design simplifies inference and enhances both temporal understanding and overall system reliability.

We posit that multimodal fusion is essential for resolving ambiguity in acoustically similar events and for enabling semantic reasoning. For example, low-frequency motor-like sounds may originate from either "vacuuming at home" or "running on a treadmill." By incorporating user habits, motion patterns, and location data, our system can differentiate these scenarios accurately. Ultimately, our goal is to produce interpretable and adaptive behavioural records to support habit tracking, self-reflection, and future human–computer interaction.

2 System Architecture Design

Our behaviour prediction system comprising three modules: the Context and Prior Knowledge Construction Module, the Audio Inference Module, and the Auxiliary Knowledge Integration Module, as shown in Figure 1.

The Context and Prior Knowledge Construction Module collects multi-modal sensor data and applies conventional inference to derive low-level context features. It also incorporates long-term user-specific priors from questionnaires, profiles, and history. The Audio Inference Module extracts acoustic features from raw audio and generates behaviour candidates based on prior knowledge. The Auxiliary Knowledge Integration Module selects and calibrates the most plausible behaviour using context cues, yielding a structured prediction.

By combining real-time context with personalised priors, the system enhances the accuracy and robustness of behaviour prediction with large language models.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

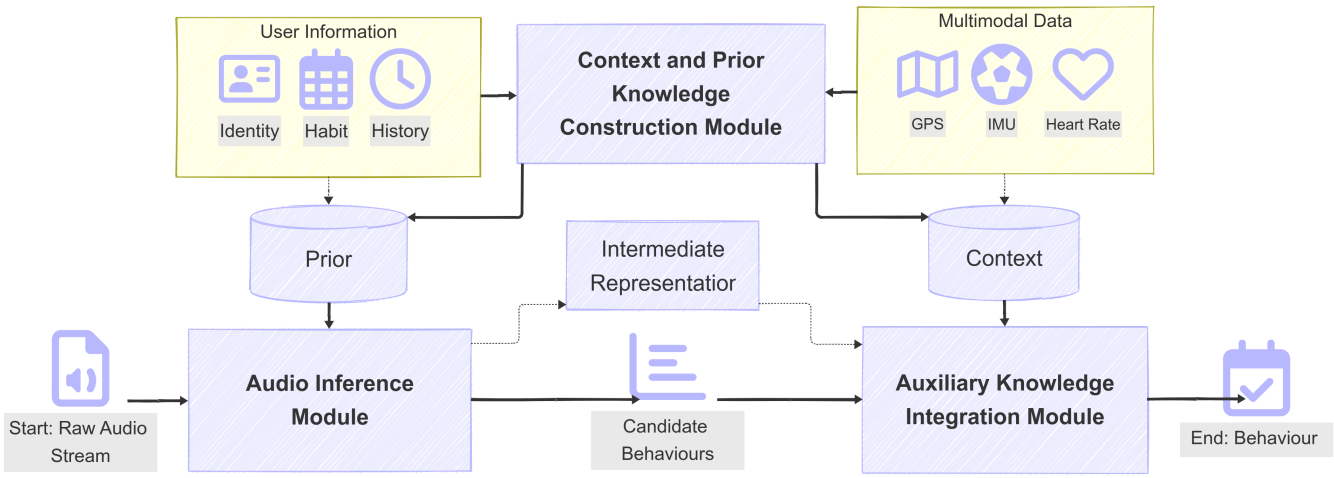

Figure 1: Overall architecture of the behaviour prediction system, consisting of three core modules: Context and Prior Knowledge Construction, Audio Inference, and Auxiliary Knowledge Integration.

2.1 Context and Prior Knowledge Construction Module

The Context and Prior Knowledge Construction Module addresses the ambiguity of audio signals by enriching inputs with structured, user-specific context. It integrates real-time multimodal sensor data with long-term personalised priors, enabling the LLM to perform semantically grounded, user-aware reasoning beyond basic acoustic event detection. The module consists of two components: Contextual Information Processing and Prior Knowledge Management, including User Profiling and Historical Behaviour Modelling.

2.1.1 Contextual Information Acquisition and Processing. Real-time context is collected from smartwatch sensors and categorised into spatio-temporal, physiological, and motion signals.

- **Spatio-temporal context**: GPS data is mapped to semantic locations (e.g., home, gym), while time information aids in inferring behaviours like resting at night or working in the morning.
- **Motion and physiological state**: IMUs (accelerometer, gyroscope) detect motion states and arm movements. Heart rate data, combined with motion, reflects activity intensity such as exercise or stress.

Raw signals are processed through a unified pipeline to produce normalised labels, as shown in Table 1.

Sensor	Raw Data	Processed Output
GPS	Latitude, Longitude	Location type (e.g., home), mobility status
IMU	Acceleration, Angular velocity	Motion state (e.g., walking), arm movement
Heart Rate	Beats per minute (BPM)	Physiological state, activity intensity

Table 1: Sensor signal processing pipeline

This unified processing ensures that multimodal signals are fed into downstream reasoning modules in a consistent, structured format, enhancing contextual coherence and inference robustness.

2.1.2 Prior Knowledge: User Profiling Module. To enable personalised behaviour recognition, the system constructs and maintains an evolving user-specific knowledge base. The profiling process comprises two phases:

(1) **Initialisation Phase**: A questionnaire captures basic demographic data (e.g., age, gender, occupation), lifestyle information (e.g., fitness habits, cafeteria usage), and routine behaviour patterns (e.g., common venues, meal timing).
(2) **Continuous Learning Phase**: As the system operates over time, it automatically analyses behavioural logs to infer and update the user's preferences and routines.

The profile structure covers three core aspects: Demographic and lifestyle attributes; Temporal behaviour patterns; Spatial-behavioural associations. An illustrative mapping between location, behaviour, acoustic, and motion cues is presented in Table 2.

2.1.3 Prior Knowledge: Historical Behaviour Module. To support long-context reasoning in large language models (LLMs) while adhering to token length constraints, we introduce a dual-buffer architecture comprising a short-term history and an hourly summary. The short-term history stores the most recent 30 behaviour records, approximately covering the past 30 minutes. In contrast, the hourly summary is updated every hour and retains the top 1 to 3 most frequent behaviour types within that period, provided their frequencies differ by no more than 20%. This hybrid memory mechanism captures both immediate changes and long-term behavioural trends, thereby enhancing temporal consistency and logical continuity in the inference process.

The user profile and buffered behavioural history are jointly encoded into a structured prior, which serves as a prompt to the audio inference module, enabling personalised and context-aware behaviour prediction.

Location	Common Behaviour	Typical Audio Clues	Motion Patterns
Home - Kitchen	Cooking	Chopping, water flow, frying	Standing, light hand motion
Office	Working	Typing, mouse clicks	Sitting, repetitive hand motion
Gym	Exercising	Impact sounds, breathing, music	High-intensity rhythmic movement
Library	Studying	Page flipping, whispering, silence	Sitting, minimal movement

Table 2: Contextualised user behaviour mapping

2.2 Audio Inference Module

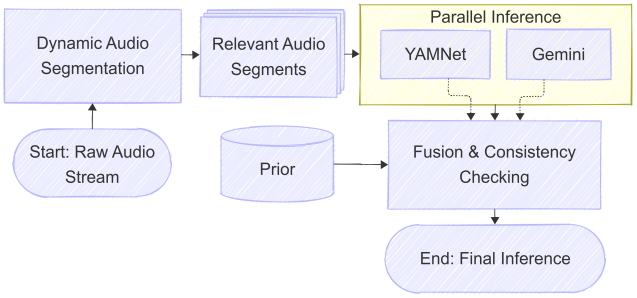

Figure 2: Architecture of the Audio Inference Module, showing dynamic segmentation, Gemini and YAMNet parallel processing, and LLM-based consistency fusion.

The audio inference module robustly infers fine-grained user behaviour segments from raw audio streams through dynamic segmentation and parallel model inference, and derives complex, coherent behaviour descriptions via consistency-driven fusion (see Figure 2. This layered architecture supports diverse acoustic conditions and enhances interpretability.

2.2.1 Dynamic Audio Segmentation. To isolate relevant audio, we apply the open-source Silero VAD, which detects voice, speech-like, and behaviour-relevant sounds (e.g., chewing, movement noise). Segmentation ensures that only content-rich clips—typically 3–10 seconds long—are passed downstream, reducing computation and improving behavioural focus.

2.2.2 Parallel Inference with Gemini and YAMNet. Each audio clip is processed by two complementary models:

- **Gemini**: A multimodal LLM that directly predicts behaviour labels from raw audio. Outputs include predicted behaviour (e.g., conversation, chewing), confidence score (0–1), and explanation based on acoustic features and context.
- **YAMNet**: A MobileNet-based model trained on AudioSet, detecting frame-level sound events (e.g., speech, impact) with associated confidence scores.

This dual-model setup improves robustness by enabling reduced model bias and misclassification, cross-validation of predictions, and resilience to noisy or ambiguous audio.

The parallel structure also supports easy integration of future modules (e.g., emotion or ambient sound recognition).

2.2.3 Fusion and Consistency Verification. In the final stage, an LLM (e.g., GPT-4o) analyses the textual outputs from Gemini and YAMNet to generate a coherent behavioural description. It consolidates segment-level reasoning and produces a list of candidate behaviours with confidence scores.

This text-based fusion simplifies architecture and leverages LLM reasoning to ensure consistency and interpretability without additional multimodal processing.

2.3 Auxiliary Knowledge Integration Module

The Auxiliary Knowledge Integration Module refines and calibrates the final behavioural prediction by leveraging contextual signals to rank and filter candidate behaviours generated by the Audio Inference Module. It enhances the system's accuracy, reliability, and interpretability through semantic consistency checks and confidence reordering.

2.3.1 Behaviour Ranking Algorithm. At the core lies the Behaviour Ranking Algorithm, which evaluates and ranks candidate behaviours based on multimodal context:

(1) **Input Processing**:
   - Receives candidate behaviours with initial confidence scores.
   - Ingests contextual information across multiple dimensions (e.g., motion, location, user preferences).

(2) **LLM-Based Scoring**:
   - *Conflict Detection*: If a candidate contradicts context (e.g., "jogging" with 30 km/h GPS), it get a score of 0.
   - *Semantic Alignment*: For valid candidates, a large language model (LLM) assigns a 1–5 score per context dimension, reflecting behavioural plausibility.

(3) **Score Aggregation**:

$$\text{Final Score} = \frac{\text{Initial Confidence} \times \sum_{i=1}^{M} \text{Context Score}_i}{5 \times M}$$

The result is normalized to $[0, 1]$.

(4) **Behaviour Selection**: The top-scoring behaviour is selected. If the score is below 0.2, the system enters fallback calibration, prompting the LLM to regenerate a prediction using intermediate representations and full context. This fallback result is flagged with a confidence of 0.

2.3.2 LLM Scoring Design Principles. To ensure robust and fair scoring from large language models (LLMs), we adopt several design principles aimed at mitigating known sources of bias. First, to address *positional bias* [13, 14], all candidate behaviours are scored independently to eliminate the influence of their order in the prompt. To reduce *length bias* [13], all behaviour descriptions are standardised in length, and the prompt explicitly instructs the model to prioritise brevity and precision. We also mitigate *format bias* [14] by applying consistent preprocessing to align all inputs with the model's expected format. Lastly, to avoid *self-enhancement*

bias [13], we adopt an ensemble scoring strategy: multiple LLMs are used in parallel, extreme scores are discarded, and the final score is computed as the average, enhancing robustness and fairness.

2.3.3 Behaviour Record Structure. As shown in Table 3, each behaviour record includes temporal, semantic, and contextual fields for structured analysis.

Field	Description
timestamp	The timestamp when the behaviour occurred
activity	The inferred behaviour label
confidence	The confidence score to this prediction
participants	People involved in the activity
location	Where the behaviour took place
description	A natural language explanation of the behaviour

Table 3: Structured Behaviour Record Format

The activity field is determined via a dynamic behaviour type pool algorithm, confidence by a ranking model, and description is auto-generated. The participants and location fields are extracted from the description using the LLM.

The behavioour type pool algorithm clusters and compresses behaviour records to maintain a dynamic set of activity types. For a new behaviour, the system uses the LLM to match its description against the existing pool. If a match is found, the behaviour is assigned to that type; otherwise, a new concise type is created and added. By avoiding reliance on fixed label sets, the system improves adaptability across diverse contexts.

3 Experimental Design

To evaluate the practical applicability of our system in real-world settings, we designed a user study to collect high-quality multimodal behavioural data for model training and validation. The study emphasises the capture of natural user behaviours while strictly adhering to privacy protection protocols.

We plan to recruit 10 participants from diverse backgrounds, all of whom are regular Apple Watch users, to cover a wide range of daily routines and lifestyle patterns. Prior to participation, each subject will sign an informed consent form and complete a questionnaire detailing demographic information, daily schedules, frequented locations, and lifestyle habits. This information serves as the foundation for constructing user-specific prior knowledge within the system.

3.1 Devices and Data Collection

Two types of devices are employed in the study:

- **Apple Watch**: We developed a watchOS application (see Figure 3) that continuously collects multimodal data from the Apple Watch, including raw audio, IMU-based motion, GPS location, and physiological signals such as heart and respiratory rates. All data streams are timestamped and synchronised from recording onset. During experiments, data are stored locally and later uploaded to the lab server for processing.
- **Wearable Camera**: Participants wear a lightweight, neck-mounted camera as shown in Figure 4 that captures daily activities from a first-person perspective. Video data are stored locally on the device.

3.2 Study Procedure

Each participant completes a full-day session to record natural behaviours. The procedure consists of the following steps:

(1) **Setup**: Participants are briefed on the study objectives and privacy policies, sign the informed consent form, complete the questionnaire, install the smartwatch app, and are equipped with the wearable camera.
(2) **Behaviour Recording**: Participants engage in their normal daily activities. A manual pause/resume function is available to suspend recording in sensitive scenarios (e.g., private conversations or bathroom use).
(3) **Data Upload**: Data from the Apple Watch are uploaded in real time or at regular intervals to the cloud. Video data are retrieved manually at the end of the session.
(4) **Device Return and Feedback**: Upon completion, participants return the camera and provide brief feedback in a structured interview regarding their user experience.

All video recordings are annotated by trained labelers to produce timestamped behaviour labels. These annotations are temporally aligned with the smartwatch data to construct a high-quality dataset for model training and evaluation. The annotated behaviours are also compared against the system's predictions to assess inference accuracy. All collected data are anonymised and processed in full compliance with data privacy and ethical guidelines.

4 Summary

Given personalised user information and access to audio and multimodal data from everyday wearable devices, our system can generate fine-grained, structured records of daily behaviour. These records serve as the basis for personalised life logging, habit tracking, and feedback delivery. To demonstrate practical value, we propose a companion application ecosystem comprising a smartwatch and mobile app, designed to provide behaviour insights and intuitive visual summaries.

Once installed and paired, the smartwatch securely uploads audio and sensor data to the cloud for anonymised processing. Inferred behaviours are then pushed to the mobile app, which offers a timeline interface where users can review their activity history by hour or drill down to minute-level details. This helps users reflect on their daily routines and identify behavioural trends.

Users may export behaviour reports over custom time ranges. The system analyses behavioural patterns and provides personalised feedback across three dimensions: Health (e.g., eating habits, activity levels, sleep), Focus (e.g., work sessions, distraction frequency), and Balance (e.g., leisure, socialising, rest). Suggestions may include increasing physical activity, reducing sedentary time, or improving focus.

By structuring behavioural data as interpretable logs, the system enhances self-awareness and supports intelligent, personalised lifestyle guidance.

References

[1] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Unb5CVPtae

[2] Qianhe Ouyang. 2023. Speech emotion detection based on MFCC and CNN-LSTM architecture. *Applied and Computational Engineering* 5, 1 (May 2023), 243–249. doi:10.54254/2755-2721/5/20230570

[3] Erasmo Purificato, Ludovico Boratto, and Ernesto William De Luca. 2024. User Modeling and User Profiling: A Comprehensive Survey. arXiv:2402.09660 [cs.AI] https://arxiv.org/abs/2402.09660

[4] David Schindler, Sascha Spors, Burcu Demiray, and Frank Krüger. 2022. Automatic Behavior Assessment from Uncontrolled Everyday Audio Recordings by Deep Learning. *Sensors* 22, 22 (2022). doi:10.3390/s22228617

[5] Johannes A. Stork, Luciano Spinello, Jens Silva, and Kai O. Arras. 2012. Audio-based human activity recognition using Non-Markovian Ensemble Voting. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. 509–514. doi:10.1109/ROMAN.2012.6343802

[6] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=14rn7HpKVk

[7] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[8] Chunzhuo Wang, T. Sunil Kumar, Walter De Raedt, Guido Camps, Hans Hallez, and Bart Vanrumste. 2024. Eating Speed Measurement Using Wrist-Worn IMU Sensors Towards Free-Living Environments. *IEEE Journal of Biomedical and Health Informatics* 28, 10 (2024), 5816–5828. doi:10.1109/JBHI.2024.3422875

[9] Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh. 2020. Speech Emotion Recognition with Dual-Sequence LSTM Architecture. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6474–6478. doi:10.1109/ICASSP40776.2020.9054629

[10] Lilin Xu, Kaiyuan Hou, and Xiaofan Jiang. 2025. Exploring the Capabilities of LLMs for IMU-based Fine-grained Human Activity Understanding. In *Proceedings of the 2nd International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things* (Irvine, CA, USA) *(FMSys)*. Association for Computing Machinery, New York, NY, USA, 13–18. doi:10.1145/3722565.3727195

[11] Ruidong Zhang, Jihai Zhang, Nitish Gade, Peng Cao, Seyun Kim, Junchi Yan, and Cheng Zhang. 2022. EatingTrak: Detecting Fine-grained Eating Moments in the Wild Using a Wrist-mounted IMU. *Proc. ACM Hum.-Comput. Interact.* 6, MHCI, Article 214 (Sept. 2022), 22 pages. doi:10.1145/3546749

[12] Xiyuan Zhang, Ranak Roy Chowdhury, Rajesh K. Gupta, and Jingbo Shang. 2024. Large language models for time series: a survey. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (Jeju, Korea) *(IJCAI '24)*. Article 921, 9 pages. doi:10.24963/ijcai.2024/921

[13] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 46595–46623. https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf

[14] Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned Large Language Models are Scalable Judges. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=xsELpEPn4A

Ethics Statement

This work investigates personalised behaviour understanding through the integration of wearable sensor data and user-specific priors. All data involved in this study—including audio, IMU, GPS, and physiological signals—are collected with informed consent and fully anonymised prior to processing. No personally identifiable information is retained or exposed at any stage.

To strictly protect user privacy, all generated behaviour records are encrypted and securely stored. These records are never accessed, shared, or used for any purpose without explicit user authorisation. Data collection is strictly opt-in, and users retain full control over their data, including the right to review or permanently delete their records at any time.

This study is committed to ethical data handling, prioritising privacy, transparency, and user autonomy throughout the entire research lifecycle.

A Experimental Device

A.1 Apple Watch App

Figure 3: Screenshots from the Apple Watch app we developed for data collection

A.2 Wearable Camera

Figure 4: A wearable camera used for capturing user behavior