EchoScriptor: Automatic Lifelogging Narratives via Activity-Based Audio–Language Model

Kaylee Yaxuan Li
Computer Science and Engineering
University of Michigan
Ann Arbor, MI, USA
yaxuanli@umich.edu

Xinghao Zhou
Computer Science and Engineering
University of Michigan
Ann Arbor, MI, USA
xinghao@umich.edu

Haizhong Zheng
Electrical and Computer Engineering
Carnegie Mellon University
Pittsburgh, PA, USA
haizhonz@andrew.cmu.edu

Kang G. Shin
Department of EECS
University of Michigan
Ann Arbor, MI, USA
kgshin@umich.edu

Alanson P. Sample
Computer Science and Engineering
University of Michigan
Ann Arbor, MI, USA
apsample@umich.edu

Abstract

Automatic, camera-free lifelogging offers new opportunities for memory rehabilitation, personal informatics, and assistive technologies. However, most existing approaches limit daily activities to isolated event labels, offering little context and lacking the narrative coherence essential for effective lifelogging. Recent advances in audio–language models combine foundation audio processing with language-based reasoning, enabling open-ended sound understanding. We introduce EchoScriptor, an end-to-end system that transforms raw in-home audio into context-aware natural-language descriptions, generating coherent narrative lifelogs of activities and acoustic contexts. In moment-level evaluation, EchoScriptor achieved 94.15% activity recognition and 89.25% background recognition accuracy, and at the summary level, achieved an F1 score of 0.92, outperforming the classifier+LLM baseline. In our user study with 20 participants across 10 household activity videos, EchoScriptor summaries were consistently rated highly, approaching the perceived utility of human-written ones. By advancing from event detection to narrative understanding, EchoScriptor establishes a significant step toward automated, unobtrusive, context-aware lifelogging technologies.

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**.

Keywords

Acoustics Sensing, Human Activity Recognition, Large Audio Language Model, Lifelogging, Memory Support

ACM Reference Format:
Kaylee Yaxuan Li, Xinghao Zhou, Haizhong Zheng, Kang G. Shin, and Alanson P. Sample. 2026. EchoScriptor: Automatic Lifelogging Narratives via Activity-Based Audio–Language Model. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3772318.3791528

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791528

1 Introduction

Enabling computing systems to passively interpret human actions and the surrounding context has long been a central goal of pervasive and ubiquitous computing research, with broad implications for health and wellness monitoring [45, 65, 102], smart environments [24, 71], assistive technologies [6, 84], and personal informatics [23, 73]. One key application is lifelogging, where daily events are automatically documented to provide contextual cues that support memory recall [58, 67]. Beyond personal reflection, lifelogging has also demonstrated as a clinical intervention for memory impairments and aging-related conditions, including dementia and Parkinson's disease, where episodic recall is critical for maintaining independence [15, 39, 89]. An effective lifelogging system should require minimal manual input, operate automatically and unobtrusively, capture rich contextual information, and present outputs in human-readable narrative form [52, 75]. In terms of sensing modalities, microphone-based sensing is promising: it is passive, already embedded in everyday devices such as voice assistants, smartphones, wearables and home appliances, and can capture a wide spectrum of ambient signals relevant to human behavior [14, 55, 62].

Human Activity Recognition (HAR), often treated as the core technology underpinning lifelogging, has primarily focused on classification, identifying which activity occurred and for how long [93]. Camera-based approaches require manual input [47, 80] and raise privacy concerns [63], while IMU- and GPS-based systems are limited to a narrow set of activities such as walking, running, or sleeping [44, 70]. More importantly,, these representations lack the expressive richness required for applications like lifelogging, personal informatics, and memory support. Natural-language descriptions therefore provide a promising direction for lifelogging. Unlike numeric labels or raw sensor streams, effective lifelogging should capture not only what activities occurred, but also how they unfolded, in what order, and within what contextual setting [5, 76, 87, 88]. In contrast, many existing systems produce large volumes of fragmented data or isolated labels without meaningful structure [16, 75].

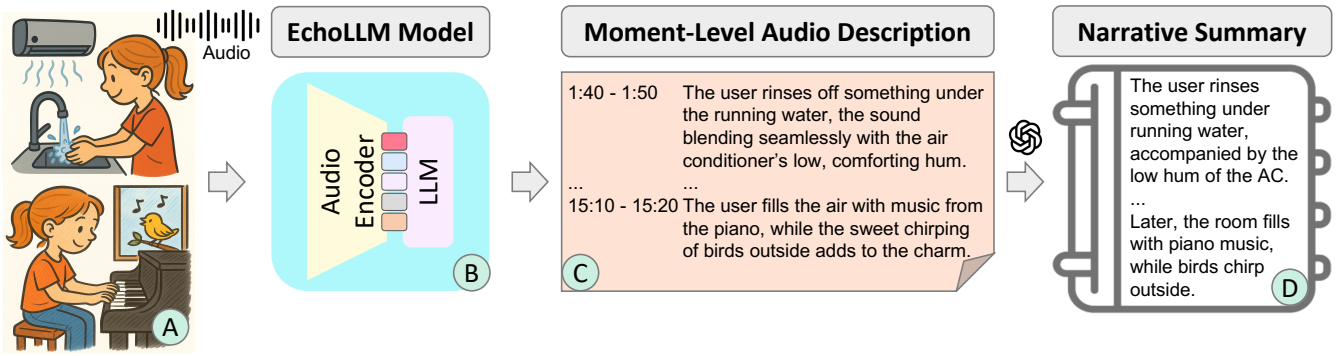

Figure 1: Overview of EchoScriptor: a generative pipeline that transforms raw household audio into narrative human activity recognition lifelogs. (A) Audio is captured from daily activities. (B) The EchoLLM human-activity large audio–language model converts raw audio to textual descriptions. (C) Moment-level audio descriptions capture user activities and acoustic context. (D) The Narrative Construction Pipeline aggregates and rewrites them into coherent, human-readable summaries.

Natural-language representations bridge this gap by shifting lifelogging from passive data collection toward context-aware, user-interpretable semantic understanding.

We propose EchoScriptor, an audio–language system that transforms raw household audio into natural-language narratives capturing both human activities and their surrounding acoustic context as shown in Figure 1. Unlike prior HAR systems that reduce audio to discrete event labels, EchoScriptor directly generates sentence-level natural-language descriptions that capture how activities and environments unfold. EchoScriptor converts raw household audio into expressive narrative lifelogs using a two-stage pipeline. First, the *EchoLLM* large audio–language model (Figure 1B) serves as the backbone for generating moment-level activity descriptions. Second, the *Narrative Summary Construction Pipeline* (Figure 1C, 1D) refines these outputs through semantic filtering, temporal aggregation, and natural-language rewriting, producing comprehensive and human-readable summaries that represent longer spans of continuous audio. Our empirical evaluation demonstrates the strong performance of EchoLLM on the first large-scale dataset of home-activity audio descriptions, comprising 199,800 synthesized audio mixtures paired with natural-language descriptions covering 24 household activities and 8 background contexts. On this benchmark, EchoLLM substantially outperformed a traditional classifier+LLM+GPT baseline, achieving 94.15% activity recognition and 89.25% background recognition. Beyond benchmark evaluation, EchoScriptor was further tested on real-world in-home recordings where it achieved an F1 score of 0.92, indicating robust generalization to naturalistic conditions. User studies further show that EchoScriptor produces summaries rated as trustworthy, useful, and nearly on par with human-written accounts, while operating significantly faster than manual annotation. Together, these contributions establish a technical and empirical foundation for lifelogging, personal informatics, and memory-support applications.

Overall, this paper makes the following contributions:

- **Audio–language system**: Introduces the first system that directly infers in-home activities with acoustic contexts from raw audio and produces coherent, context-aware natural-language narratives, with an interactive website front-end accessible from any microphone-enabled device.
- **Empirical validation:** Developed EchoLLM and demonstrated its strong performance with the first large-scale home-activity audio description dataset, substantially outperforming the classifier+LLM+GPT baseline.
- **User and real-world studies:** Demonstrates that EchoScriptor generates summaries rated as trustworthy, useful, and nearly on par with human written accounts, while operating much faster than manual annotation.

2 Related Work

This section reviews prior work in three areas: audio-based human activity recognition (HAR), lifelogging for memory support, and the integration of LLMs with sensor data. It highlights their advantages and limitations, and identifies the gaps that motivate this work.

2.1 Audio-based Human Activity Recognition

Human Activity Recognition (HAR) has been widely studied in mobile systems, wearable technology, healthcare, and ubiquitous computing, with modalities such as cameras [57, 81], IMUs [7, 43, 95, 96, 98], microphones [54, 62], GPS [30, 94], and RF signals [13, 51], each presenting distinct trade-offs in coverage, reliability, and practicality. Of these, audio-based sensing stands out for its low power cost, unobtrusive nature, and minimal instrumentation requirements [56]. Unlike vision- or RF-based systems, microphones are less affected by occlusions or line-of-sight constraints, making them versatile for everyday environments. Audio-based systems, however, face two practical challenges: recordings may capture speech, raising privacy concerns, and signals are susceptible to ambient noise. To mitigate these problems, researchers have explored ultrasonic sensing to capture signals beyond speech frequencies [42, 64] and surface acoustic waves to reject environmental noise for more robust recognition [29, 60].

Most audio-based HAR systems follow a closed-set classification framework, in which short audio segments are assigned labels from a predefined taxonomy. This reliance on supervised learning

limits scalability, as it is infeasible to enumerate or collect training data for the vast range of human activities and all possible combinations [21, 101]. Fixed closed-set class definitions and labeling requirements also introduce high annotation costs, while adapting to new environments often necessitates retraining [32, 72, 91]. Although effective for constrained tasks, this formulation struggles to capture overlapping or concurrent events [1, 69] and frequently ignores contextual cues such as ambient sounds. Finally, the outputs are restricted to discrete labels, which limits expressiveness and prevents the generation of temporally grounded, context-rich accounts of daily activities [85, 94]. These limitations underscore the need to move beyond closed-set classification toward structured, interpretable, and context-aware representations of human activity. EchoScriptor directly addresses this need by generating natural-language narratives that capture activities together with their acoustic contexts, providing richer and more expressive outputs than discrete labels can offer.

2.2 Lifelogging Systems and Memory Support

The need for such representations is particularly acute in lifelogging, where the value of the system depends on transforming raw sensor data into meaningful accounts of daily life that support memory and reflection. Effective lifelogging depends on automated understanding of human activities. Existing systems either rely on wearable cameras that require extensive manual annotation [58, 77], or employ passive sensors that generate large volumes of raw, disconnected data without meaningful interpretation [38, 46]. The absence of automated, human-readable summarization therefore severely limits the practical utility and clinical effectiveness of lifelogging, constraining downstream applications such as memory support, Alzheimer's treatment, and elderly care. Advancing lifelogging requires HAR algorithms capable of automatically producing coherent, human-readable summaries. Such algorithms must capture detailed context continuously with minimal user burden or intervention, thereby improving the interpretability and usability of lifelogs [90]. Among sensing modalities, audio is particularly promising due to its passive, unobtrusive nature and its ability to provide rich contextual cues from everyday environments [4, 53, 62]. Although audio has been incorporated into lifelogging, prior work has largely reduced it to discrete activity labels [18, 78, 79] or treated it as a complementary modality alongside vision, user input, WiFi, or GPS [3, 82]. To date, no audio-only lifelogging system has generated narrative, language-based outputs directly for end users. EchoScriptor fills this gap by producing coherent, human-readable summaries from raw household audio, making lifelogs more interpretable, usable, and effective.

2.3 Large Audio Language Model (LALM) and multimodal LLM

These limitations motivate approaches that can directly generate expressive natural language descriptions from audio. Recent advances in large audio language models (LALMs) have shown the ability to map complex acoustic input to fluent textual outputs [34, 36, 49, 83]. However, existing LALMs primarily target general audio understanding and have not been designed to capture human activities or daily contexts. Recent efforts have explored integrating large language models (LLMs) with multimodal sensor data for activity interpretation and journaling. For example, HARGPT [43] applies LLMs to interpret IMU signals, while SensorLM [100] extends this idea to encode physiological and motion data such as heart rate and step counts. In general, these projects focus primarily on the technical challenge of representing sensor streams within LLMs to achieve accurate recognition [12, 26]. Others adopt a simpler pipeline in which sensor-derived labels are passed to an LLM (e.g., ChatGPT, Gemini) to generate textual descriptions [41, 85]. Beyond technical modeling, some systems demonstrate how sensor–LLM integration can enable user-facing applications. AutoLife [94] combines GPS traces, environmental context, and videos to construct daily locomotion journals. LLMSense [68] employs a two-stage pipeline in which a neural network infers activities and timelines; these outputs are then passed to an LLM for higher-level reasoning such as inferring cognitive states or occupational patterns. MindScape [66] is a mobile application that collects self-reports on daily experiences and uses LLMs to generate reflective mental-health journals. Overall, prior systems either concentrate on encoding low-level sensor data for classification or rely on multimodal fusion and user input. None of these explicitly focuses on continuously modeling daily home activities optimized for lifelogging applications with user-centered output. To bridge this gap, we introduce EchoScriptor, an audio-only system designed to generate coherent, narrative-style summaries that capture both activities and their surrounding acoustic context, laying the foundation for more accessible, context-aware, and user-centered lifelogging.

3 System Design: EchoScriptor Overview

This section details the design of EchoScriptor, which integrates two stages: the EchoLLM model and a post-processing pipeline that converts model outputs into user-facing lifelogs. EchoScriptor operates directly on raw audio from arbitrary devices without device-specific calibration or heavy preprocessing. In Stage 1, EchoLLM generates **moment-level descriptions**, short natural-language descriptions that characterize activities and environmental sounds within 10-second audio segments. The post-processing pipeline Stage 2 then consolidates these descriptions into **narrative summaries**, structured activity logs and fluent, chronologically ordered accounts spanning an extended recording. Together, moment-level descriptions and narrative summaries provide both fine-grained evidence of recognition and coherent accounts of daily activities suitable for lifelogging applications.

3.1 Stage 1 — EchoLLM: Large Audio Language Model for Moment-Level Audio Descriptions

EchoLLM builds on a large audio–language modeling framework that integrates pretrained representations from both modalities. As shown in Figure 2, the system consists of three main components: an audio encoder (CAV-MAE) that converts raw 16 kHz audio into high-dimensional embeddings, a text encoder that embeds the instruction prompt to guide structured understanding, and a projection module that aligns the two modalities in a shared representation space. The aligned features are then passed to a large language model, which generates natural-language descriptions of the audio input. The following subsections detail each component.

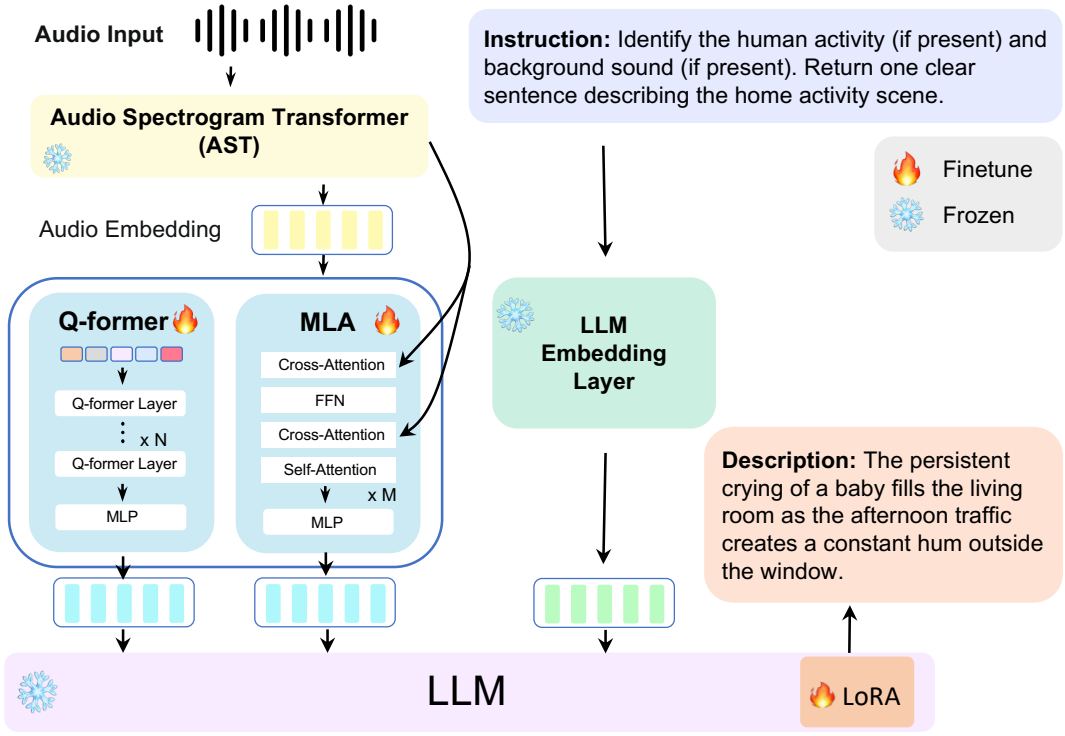

Figure 2: Architecture of EchoLLM for moment-level audio descriptions. Raw audio is encoded by the Audio Spectrogram Transformer (AST), adapted through a Q-Former and multi-layer aggregator (MLA) to align with the LLM embedding space, and combined with textual instructions. The LLM, fine-tuned with LoRA, generates natural-language descriptions of activities and acoustic context.

3.1.1 Audio Encoder. The audio encoder transforms raw audio signals into high-level representations suitable for subsequent language modeling. The Audio Spectrogram Transformer (AST) [35], a widely used Transformer-based architecture for audio embedding extraction, is adopted as the backbone. The audio encoder converts input audio into hidden representations through self-attention mechanisms. This pre-trained encoder has demonstrated strong performance across diverse audio classification and event detection tasks, making it a suitable backbone for downstream language modeling [37, 50].

In this paper, the audio encoder (Audio Spectrogram Transformer) from CAV-MAE [37] is used, instantiated as a 12-layer Transformer (hidden size 768, 12 attention heads, ≈85M parameters) pretrained in a self-supervised masked autoencoding regime on AudioSet-2M [31] and VGGSound [11]. Specifically, each 10-second segment is converted into 128-bin log-Mel filterbank features using a 25 ms Hamming window with a 10 ms hop, yielding a 1024×128 time-frequency spectrogram. The spectrogram is split into 512 overlapping 16×16 patches, each linearly projected to a 768-dimensional embedding with learned positional encodings, giving the model a global receptive field from the lowest layers. The pre-training objective randomly masks a large fraction of the spectrogram patches and reconstructs the missing content, producing robust and semantically meaningful representations. Since the goal is feature extraction, the classification head is removed and last-layer token features are taken as AST outputs. The input interface and pre-training choices above standardize the encoder across devices and recording conditions, while the fixed 10-second tokenization (512 patches) provides a stable sequence length for efficient batching. During training in EchoScriptor, all AST parameters are kept frozen to preserve the generality of the pre-trained encoder and to focus optimization on the modality adapter and the language model.

3.1.2 Modality Adapters. Unlike textual tokens, the embeddings produced by the audio encoder cannot be directly consumed by the language model. This mismatch arises because the audio encoder operates in its own representational space, which differs from the token embedding space of the language model. To bridge this gap, modality adapters are introduced following GAMA [33]. The adapters consist of two key components: an audio Q-Former and a multi-layer aggregator.

Audio Q-Former. The Q-Former is based on the Querying Transformer architecture [59], initialized from BERT [19] with a set of learnable query tokens. It takes as input the final-layer features from the AST and projects them into a semantically rich and generalized representation space. Trained with a combination of audio–text matching, grounded text generation, and contrastive objectives, the Q-Former captures high-level semantic information that aligns audio features with the language model's embedding space. In the

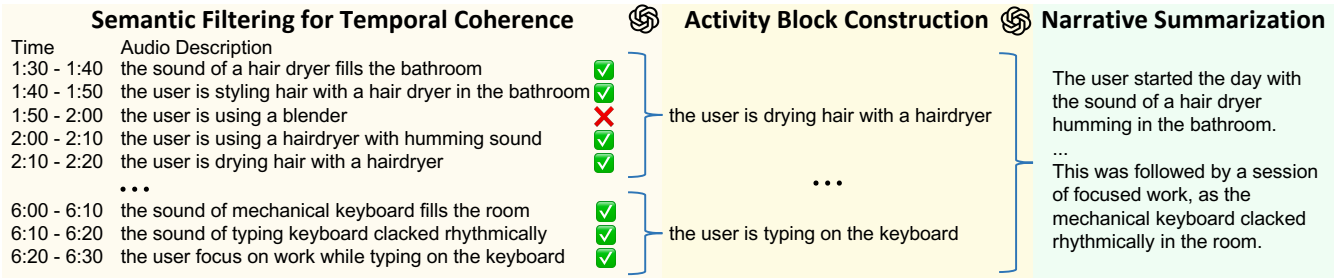

Figure 3: Narrative Summary Construction Pipeline of EchoScriptor: (1) Semantic filtering enforces temporal coherence by removing brief anomalies, (2) Activity block construction consolidates consecutive events into higher-level representations, and (3) Narrative summarization converts activity blocks into coherent, human-centered lifelogs.

EchoLLM model, the Audio Q-Former is initialized from the pre-trained checkpoints provided by GAMA [33] and further fine-tuned on the task-specific dataset.

Multi-layer Aggregator. While the Audio Q-Former relies only on the final-layer features of AST, the multi-layer aggregator integrates representations from multiple intermediate layers together with the final-layer outputs. Different layers of AST capture auditory information at varying levels of abstraction, from low-level spectral textures to high-level acoustic concepts. The aggregator fuses these signals using a Transformer-style module with self-attention and cross-attention, producing a holistic and fine-grained representation of the input audio. This representation further strengthens cross-modal alignment with the language model. As with the Audio Q-Former, the multi-layer aggregator is initialized from pre-trained GAMA [33] weights and further fine-tuned on the task-specific dataset.

3.1.3 Autoregressive Large Language Model. The final component of EchoLLM is an autoregressive large language model that generates natural-language responses conditioned on the adapted audio features. Specifically, EchoLLM employs LLaMA 2-7B [86], a decoder-only Transformer pretrained on large-scale text corpora. Given the projected audio embeddings from the modality adapters together with textual instructions, the language model autoregressively predicts the next token, producing coherent and contextually grounded descriptions of the audio content.

3.2 Stage 2 — Narrative Summary Construction: From Moment-Level Audio Descriptions to Human-Centered Activity Summaries

The EchoLLM model produces natural-language descriptions for every 10-second audio segment, yielding fine-grained moment-to-moment acoustic context. While these outputs provide detailed system-level evidence of recognition, they are not directly suitable for lifelogging applications, where users expect concise, coherent, and human-readable narratives of daily activities. To bridge this gap, EchoScriptor incorporates a post-processing pipeline that transforms raw segment-level captions into structured activity logs and fluent diary-style summaries. This pipeline consists of three stages: rule-based filtering to reduce noise and anomalies, temporal aggregation into coherent activity blocks, and natural-language generation of user-centered summaries. Figure 3 illustrates an example. The following elaborates on the detailed design.

(1) **Semantic Filtering for Temporal Coherence.** Since raw segment-level predictions may contain noisy or misclassified events, a filtering step is required to enforce temporal and contextual consistency. The key motivation is context awareness: human activities are typically continuous, and brief anomalies (e.g., a one-off "chopping" event during a sustained hair-dryer sequence) rarely represent true events. While conventional numerical signals can be smoothed with low-pass filters, text-based activity descriptions demand semantic filtering. To this end, GPT-4o was prompted with explicit rules: (i) continuous sounds (e.g., hair dryer, mechanical keyboard, chopping) were retained only if they appeared in at least three consecutive segments; (ii) faint background sounds (e.g., fridge hum, air conditioner) were retained only if they persisted across four or more consecutive segments; and (iii) short activities (e.g., toothbrush, talking) were retained only when they occurred at least twice consecutively. These rules enforce temporal coherence and ensure that isolated errors and brief anomalies are excluded from the lifelogging output.

(2) **Activity Block Construction.** After the initial filtering, a second layer of semantic processing was applied to consolidate the remaining sequences into coherent activity blocks. This step leveraged GPT-4o to enforce temporal and spatial continuity, ensuring that related events were grouped together into higher-level representations. For example, consecutive "hair dryer in the bathroom" segments were merged into a single block, whereas brief anomalies (e.g., a mislabeled "chopping" slice within the sequence) were identified as inconsistent events and removed.

(3) **Narrative Summarization into Human-Centered Lifelogs.** The cleaned activity blocks were transformed into a fluent diary-style narrative. GPT-4o was prompted to rewrite the sequence into a concise third-person account This stage preserves the chronological order of events and generates a coherent paragraph-style lifelog suitable for memory support applications.

Taken together, the EchoLLM model and the post-processing pipeline form EchoScriptor, the first two-stage audio-based system capable of generating temporally structured and context-aware

activity narratives. Unlike prior approaches that rely on discrete activity labels or handcrafted lifelogging rules, EchoScriptor directly transforms raw audio into natural-language summaries that capture both primary activities and background context. This design bridges low-level acoustic perception with high-level narrative representation, enabling lifelogs that are not only technically accurate but also coherent and user-centered.

4 Moment-Level Descriptions Evaluation Setup

This section presents the experimental setup for evaluating EchoLLM on moment-level activity recognition tasks, where each short audio segment is paired with a natural-language description of the user's activity and the accompanying environmental sound. We first describe the dataset design and training setup and then introduce the baseline approach, which combines an audio classification model with an LLM. We conducted evaluation using two complementary strategies: (i) LLM-based scoring to approximate human judgments, and (ii) standard NLP metrics for sentence- and semantic-level similarity. Together, these metrics assess both factual accuracy and narrative coherence of audio descriptions.

4.1 Dataset Design for Moment-Level Descriptions

Current audio datasets primarily provide categorical labels, lacking detailed descriptions of activities and their surrounding context. However, human-readable narratives that capture foreground user interactions and ambient environments will improve daily-activity logging's interpretability, effectiveness, and practical utility [8, 52]. Although datasets such as AudioCaps [48] and Clotho [20] provide natural-language captions for general purposes, they are not designed to represent human daily home activities. To address this gap, we introduce, to our best knowledge, the first audio-description dataset focused on daily home activities. It is specifically designed for home-activity logging and provides natural-language descriptions that capture both primary activities and associated environmental sounds.

The dataset centers on realistic home scenarios, each paired with detailed natural-language descriptions. Initially, 740 authentic home scenarios were generated using ChatGPT-4o. Each scenario was explicitly defined by three components: (1) a foreground user activity (e.g., coffee grinding), (2) an associated ambient sound (e.g., air conditioning), and (3) a time-of-day tag (morning, afternoon, evening, or night). The design spans 24 kinds of household human activities and 8 common home background noises listed in Table 1. To acoustically instantiate these scenarios, 10—13 long-form raw recordings of each activity and background, typically ranging from several minutes to about one hour, were collected from public sources such as YouTube and Freesound. These sources offer broad coverage and natural variability, making them well suited for capturing diverse household conditions. From each long recording, three non-overlapping 10-second clips were extracted, yielding 30–39 clips per type of sound. Whenever possible, longer recordings were selected so that contained natural variations over time, ensuring that 10-second clips are sampled from different intervals differ substantially. For short-duration sounds (e.g., toilet flushing), we instead gathered 30 distinct 10-second clips to ensure intra-class diversity. This sampling strategy avoids excessive redundancy from long recordings and maintains balanced class sizes. Clips were partitioned into distinct training and testing pools to rigorously avoid data leakage. Activities used a 9:1 split (≈27 training, 3 testing clips per class), while backgrounds adopted an adjusted split (≈10 training, 6–8 testing clips per class) to increase diversity in the test set, which is critical for evaluating generalization under varied ambient conditions.

For each home scenario, audio mixtures were synthesized by overlaying every foreground activity clip from the training pool with every background clip from the corresponding pool. Foreground activities were mixed at approximately +5dB relative to the background, ensuring that the activity remained perceptually dominant while preserving the surrounding ambient context. For example, combining the "coffee grinder" activity (≈27 training clips) with the "air conditioning" background (≈10 training clips) yields ≈270 distinct mixes within that activity-background scenario. Extending this combinatorial pairing across all 740 scenarios yields 199,800 synthesized training clips. Thus, despite the modest size of the raw pools, the cross-product expansion ($27 \times 10 \times 740 = 199{,}800$) generates a large and diverse training set.

Each synthesized audio instance was paired with natural-language descriptions generated by ChatGPT-4o. Prompts were conditioned on the scenario definition (user-activity, background, and time-of-day) and explicitly constrained to mention only audible elements, avoiding unverifiable details such as emotions or intentions. For each scenario, we generated 90 natural-language descriptions to introduce lexical and stylistic diversity. These descriptions were paired with the corresponding audio mixtures from the same scenario and manually verified to ensure semantic alignment and eliminate hallucinated content. Because each scenario contained multiple audio mixtures, a single description was associated with multiple audio instances within that scenario. All pairings underwent a manual verification step to confirm validity, ensuring reliable (audio, description) pairs for both training and evaluation. More specifically, it is to ensure that the synthesized descriptions remain faithful to the underlying activity recordings. Prior to verification, substantial effort was invested in prompt design to minimize hallucinations and ensure the description is accurate in the generated descriptions. The final prompt configuration effectively constrained the model to avoid introducing hallucinated elements such as imagined events, locations, additional people, emotions, or speech content. Three researchers conducted the verification. For each scenario, approximately one-third of the descriptions were reviewed in full, covering different lexical variants and writing styles. The remaining descriptions underwent a targeted consistency check, focusing on identifying unsupported acoustic events, implausible details, or attributes that contradicted the scenario specification. Descriptions containing hallucinated or extraneous content were corrected when feasible or removed directly. Overall, fewer than 5% of the dataset required modification or removal, a low error rate attributable to the rigorous iterative prompt engineering employed prior to generation. Although formal inter-annotator agreement was not computed, the reviewers first jointly inspected a subset of scenarios to calibrate acceptance criteria. Any ambiguous cases encountered during individual review were resolved through consensus. This process ensured that all retained descriptions were

Table 1: Overview of the proposed dataset, including representative household activities, background sounds, and detailed training dataset statistics.

Household Activities			Background Sounds	
Shower	Electric Razor	Electric Toothbrush	Air Conditioning	Traffic Noise
People Talking	Toilet Flushing	Hairdryer	Birds Chirping	Raining
Petting Dog (Barking)	Playing Video Games (Mario)	Baby Crying	Fridge Humming	Laundry
Petting Cat (Meowing)	Playing Piano	Typing on Keyboard	Dishwasher	Footsteps
Printer	Phone Ringing	Shredder		
Blender	Microwave	Coffee Grinder		
Food Processor	Chopping Board	Running Water in Sink		
Timer Beeping	Ice Maker	Vacuum Cleaner		

Category	Count	Notes
Raw Long Activity Recordings	313	Household human activity sounds from online sources
Raw Long background Recordings	80	Environmental background sounds from online sources
Train activity clips	648	Moment-level 10s audio clips for each activity
Train Background clips	80	Moment-level 10s audio clips for each background
Synthesized training scenarios	740	Authentic home scenarios
Synthesized training audio clips	199,800	Activity and background mixtures for audio scenes of every scenario
Natural-language audio descriptions	66,600	Textual descriptions of audio scene for activity and background
Pure activity training audio clips	1,077	Pure activity clips before augmentation
Pure background training audio clips	1,956	Pure background clips before augmentation
White noise training clips	4,382	Artificial noise samples representing null/no-event conditions
Total training audio clips	**463,532**	All training data (mixtures + augmented pure clips + white noise)

semantically aligned with the corresponding audio mixtures and that verification decisions were consistent across annotators.

In addition to the mixed activity–background combinations, we constructed pure activity, pure background, and white noise conditions. The raw pools contained 1,077 pure activity clips and 1,956 pure background clips. To ensure comparability with the overlapping mixtures, these recordings were expanded into large training sets using a two-step procedure. First, long-form recordings were partitioned into multiple non-overlapping 10-second segments, capturing natural acoustic variations across time. Second, each segment was augmented with transformations, including volume scaling and frequency equalization, simulating everyday variability such as different loudness levels, timbre changes, and recording conditions. This process yielded 161,550 pure activity clips and 97,800 pure background clips in the training set, each paired with natural-language descriptions. White noise clips were also introduced to represent null conditions with no meaningful activity or background, producing 4,382 for training and 486 for testing. Incorporating these single-source and null conditions broadens coverage to simpler but realistic scenarios, strengthens model robustness, and provides a reference for distinguishing meaningful events from silence or irrelevant input.

An additional evaluation set of 192 new scenarios was constructed exclusively from audio clips reserved for testing, ensuring no overlap with the training pool. This design ensures strict source separation and prevents cross-scenario leakage, enabling unbiased evaluation of model generalization. For each scenario, 10 new natural-language descriptions were generated and manually verified, yielding 1,920 test descriptions in total. The same activity–background mixing strategy used in training was applied to create test mixtures, and the test set also includes pure activity, pure background, and white noise conditions to support evaluation across single-source and null cases. Compared to the training set, the test set is smaller because the training set is inflated by combinatorial mixing (e.g., many activity–background combinations), whereas the test set is drawn directly from held-out raw clips and emphasizes diversity rather than scale. This design provides sufficient coverage for evaluation, while the more meaningful and practical assessment of generalization will come from future real-world data collection.

4.2 Training Procedure for EchoLLM

For EchoLLM, training is initialized from the final checkpoint of GAMA, which serves as the pretrained backbone. As detailed in Section 3.1, GAMA undergoes a two-stage training process: it is first fine-tuned on large-scale audio–language datasets (e.g., OpenAQA [36], MusicCaps [2], NSynth [22]) to acquire perceptual abilities for audio event recognition, and is then instruction-tuned over CompA-R [33] dataset to develop multi-step reasoning about acoustic scenes. While this process equips GAMA with broad perceptual coverage and contextual reasoning ability, direct application to the proposed dataset yielded unsatisfactory results. With the evaluation metrics described in Section 5.1.1, only 9.74% of the test samples received a score of 5, corresponding to a fully correct activity description. Manual inspection further confirmed that the generated descriptions were overwhelmingly inaccurate in practice. The poor performance highlights GAMA's limitations in home activity scenarios, particularly with respect to human-centered sounds and

overlapping events, thereby motivating the need for further fine-tuning. As labeled in Figure 2, only certain modules are finetuned. Each audio sample is formulated as a question–answer pair, where the question asks what the user is doing or what is happening, and the answer provides the corresponding audio description. Since the question is identical across all pairs, the original loss function, which incorporated both question and answer, was modified by masking out the question component. This adjustment ensures that optimization focuses solely on the answer, thereby improving the quality of generated audio descriptions. The masked cross-entropy loss is defined as:

$$\mathcal{L}_{\text{masked}} = -\sum_{t=1}^{T} m_t \log \hat{y}_{t,y_t}, \tag{1}$$

where $\hat{y}_{t,y_t}$ is the predicted probability of the ground-truth token y_t at step t, and $m_t \in \{0,1\}$ is a mask that excludes the question tokens ($m_t = 0$) and keeps only the answer tokens ($m_t = 1$). In our implementation, masked tokens are assigned the ignore index -100, which PyTorch's `CrossEntropyLoss` omits from optimization. The model is fine-tuned for five epochs on four NVIDIA RTX 6000 Ada GPUs. For parameter-efficient adaptation, Low-Rank Adaptation (LoRA) [40] is applied to the LLM training (rank $r = 8$, $\alpha = 16$, dropout $p = 0.05$), enabling the fine-tuning process to preserve knowledge acquired during the original GAMA training, while substantially reducing GPU memory usage and computational cost.

4.3 Baseline: AST (classifier) + LLM Pipeline

EchoScriptor proposes an integrated pipeline that generates human activity lifelogging summaries directly from audio input, producing richer contextual representations through natural-language descriptions. As a comparison baseline, a conventional machine learning approach is considered, in which an audio-based activity classification model first predicts activity labels, and the predicted labels are then provided to a large language model to generate lifelogging descriptions. More specifically, the classification stage employs the Audio Spectrogram Transformer (AST) [35], a model that has achieved strong performance on large-scale audio classification benchmarks. AST is a convolution-free, attention-only architecture. Input audio is first converted into spectrogram images, which are then divided into overlapping 16×16 patches, each linearly projected into a one-dimensional embedding. These embeddings are processed by a Transformer pretrained on ImageNet [17], effectively transferring visual feature representations to audio and enabling AST's strong performance on classification tasks. For a fair comparison, the baseline employs the same large language model (LLaMA-2-7B) used in the EchoLLM framework, generating descriptions from the labels predicted by AST. Evaluating this label-based pipeline against the proposed integrated framework provides a rigorous assessment of the benefits and performance gains of direct audio-to-summary generation. The AST model was originally trained on AudioSet, which contains 527 sound classes. Among the classes in our dataset, 27/34 classes of audio overlap with those included in AudioSet. To evaluate its applicability, inference was first done directly on the test set without additional fine-tuning. The test data corresponds to the activity–background overlapping subset described in Section 4.1. For evaluation, each overlapping audio sample was assigned both an activity label and a background label as ground truth. During inference, the model predicted the two most probable labels, which were then interpreted as the activity and background labels, regardless of order. The results show that for the activity classes in our dataset that overlapped with AudioSet, the correct activity appeared within the top two predictions for 39.13% of the samples, the correct background appeared for 19.77%, and both activity and background were simultaneously identified in only 2.44%. This poor performance indicates that the AST classification model, when applied out of the box, does not generalize to overlapping home activity scenarios and therefore requires fine-tuning. To address this limitation, the model was fine-tuned on our dataset for five epochs, with cross-validation used to select the best-performing checkpoint. Using the same top-2, order-agnostic accuracy calculation, performance improved substantially: the correct activity and background were simultaneously identified in 91.85% of the samples, and in 72.85% of the samples when the order of predictions was considered. After label prediction by the classification model, the LLM was prompted to generate realistic and contextually appropriate descriptions of home activities. For fair comparison, the EchoLLM model and the AST+LLM baseline serve as aligned counterparts for moment-level descriptions. The baseline uses a LLaMA-based language model because EchoLLM is likewise built on the same LLaMA backbone, ensuring architectural parity at this stage. EchoLLM+GPT (i.e., EchoScriptor) and AST+LLM+GPT constitute the corresponding configurations for the narrative-summarization stage. Their comparative performance will be analyzed in detail in Section 6.

5 Moment-Level Evaluation Results

This section evaluates the capabilities of EchoLLM, the component of EchoScriptor responsible for generating moment-level descriptions, in comparison with the baseline system AST+LLM. The objective is to assess whether the model can accurately characterize acoustic scenarios involving human activity in domestic environments. The evaluation considers four representative conditions in which clips: (i) contain only human activity, (ii) contain only background sounds, (iii) contain simultaneous activity and background sounds, and (iv) contain no meaningful event. These conditions reflect natural variations in everyday settings and are essential for assessing practical utility. The test set was partitioned accordingly into four categories: activity–background overlap (4,920 clips), pure background (213 clips), pure activity (106 clips), and noise-only with no meaningful event (486 clips). The construction of the overlap dataset is described in Section 4.1. All clips were paired with the same instruction prompt to ensure consistency across conditions: "*You will be provided with an audio segment recorded at home. The audio may contain only user activity sounds, only background sounds, both overlapping, or none. Identify the human activity (if present) and background sound (if present), and return one clear sentence describing the home activity scene.*" The ground truth for evaluation corresponds to the reference audio descriptions defined in Section 4.1, which capture the annotated scene for each clip. The evaluation proceeds in two stages. First, the capability of EchoLLM to automatically distinguish among four representative conditions is assessed: activity-only, background-only, overlapping activity and

Table 2: Recognition accuracy on pure subsets (106 activity-only clips, 213 background-only clips). Source–type detection indicates correct classification of clip type, and description accuracy indicates whether the predicted content matches the reference.

Source-type Category	Source–type Detection	Source–type Detection Error	Description Accuracy
Pure Activity	94.34%	5.66%	99.06%
Pure Background	100.00%	0.00%	98.13%
Weighted Average	**98.12%**	**1.88%**	**98.43%**

background, and noise-only. These categories capture the prevailing patterns of household audio and are critical for deployment, as domestic recordings naturally alternate between them. In contrast, the AST+LLM baseline cannot do this discrimination: it outputs only a ranked list of class probabilities and requires prior specification of how many prediction labels should be interpreted as present. Although more complex overlaps, such as multiple simultaneous activities or backgrounds, remain an important direction for future work, focusing on these four conditions provides a balanced trade-off between ecological validity and tractable evaluation in practical household environments.

5.1 Evaluation Protocols

Two complementary protocols were employed to assess the quality of generated descriptions. The first uses an LLM judge to provide human-like evaluation of activity accuracy, background accuracy, and overall semantic expression. The second applies established NLP similarity metrics to quantify sentence- and embedding-level alignment with the ground-truth descriptions. Together, these protocols capture both subjective interpretability and objective semantic correspondence, providing a comprehensive evaluation framework.

5.1.1 Human-Like Evaluation via LLM Judge. Recent studies have shown that large language models (LLMs) can function as reliable judges for both language and multimodal evaluation tasks. Following this paradigm, ChatGPT-4o was employed to assess the quality of generated audio descriptions under a single, fixed instruction. The evaluation considered three criteria: (i) *human activity accuracy*, (ii) *background sound accuracy*, and (iii) *overall semantic expression quality*, defined as the global alignment of the predicted sentence with the reference in terms of scene, setting, and coherence. Each criterion was scored on a 0–5 scale by comparing the predicted sentence with the ground-truth description in the context of everyday home activities and background sounds. A score of 5 indicates identical or synonymous meaning (e.g., "vacuuming the living room" vs. "using a vacuum cleaner"). A score of 4 reflects a closely related or strongly implied meaning (e.g., "taking a shower" vs. "water running in the bathroom"). A score of 3 denotes that the prediction and reference belong to the same broader category but differ in the specific action (e.g., "washing face" vs. "taking a shower"). A score of 2 captures a loose relation with limited semantic overlap (e.g., "using a microwave" vs. "making coffee"). A score of 1 represents minimal relation, such as two different grooming activities in the same setting (e.g., "using an electric toothbrush" vs. "using an electric shaver"). A score of 0 is assigned when the prediction and reference are unrelated (e.g., "using a hair dryer" vs. "vacuuming"). If a caption omitted either human activity or background content, the corresponding criterion was marked as `N/A`, while the remaining criteria were still evaluated. All `N/A` assignments were manually verified for correctness, with verification restricted to confirming omissions without altering any scores produced by ChatGPT-4o.

5.1.2 Semantic Similarity Metrics. To complement label-level accuracy, the quality of generated descriptions was also assessed at the sentence and semantic levels. Two widely used natural-language processing metrics were employed: BERTScore-F1 [28, 61, 99], which measures semantic similarity between generated and reference texts using contextual embeddings, and Sentence-BERT [27, 74, 97], which evaluates sentence-level alignment through cosine similarity in the embedding space.

5.2 Results - Recognition Across Acoustic Conditions

Reliable discrimination and context-aware identification of user activity are essential to avoid false detections in practical household environments and to provide accurate interpretation when activities and background sounds interact. This subsection evaluates EchoLLM under these fundamental conditions to establish its reliability.

5.2.1 Recognition on Pure Activity and Background Subsets. For the activity–background overlap subset, the model consistently detected the presence of both activity and background, achieving 100% recognition of overlapping conditions. To further examine performance in simpler cases, analysis was conducted on the pure subsets containing only activity or only background sounds. As shown in Table 2, the model reached 98.12% accuracy in source-type detection and 98.43% accuracy in description correctness for both pure activity and pure background clips. These results highlight two key findings. First, the model can reliably distinguish activity-only and background-only conditions from overlapping cases, a capability unavailable to traditional classification models. Second, EchoLLM demonstrates high reliability in identifying the specific activity or background event in single-source scenarios, establishing a strong foundation for evaluation under more complex conditions.

5.2.2 Recognition on Noise-Only Subsets. The evaluation further included 486 clips containing only background white noise, each paired with the ground-truth description "There is no meaningful home event happening." EchoLLM correctly identified all cases,

Table 3: Comparison of AST+LLM and EchoLLM under LLM-as-a-Judge evaluation. Mean scores are on a 0–5 scale; proportions indicate the percentage of clips with scores ≥ 4 or = 5.

	Human Activity Accuracy			Background Accuracy			Overall Expression Quality		
Model	Mean	Prop. ≥ 4	Prop. = 5	Mean	Prop. ≥ 4	Prop. = 5	Mean	Prop. ≥ 4	Prop. = 5
AST+LLM	4.69	93.48%	86.72%	4.50	86.48%	83.93%	4.22	75.72%	64.89%
EchoLLM	**4.92**	**98.92%**	**94.15%**	**4.64**	**89.74%**	**89.25%**	**4.72**	**91.38%**	**85.45%**

achieving 100% accuracy. This result demonstrates that the model does not hallucinate activities in the absence of meaningful events, underscoring its robustness and reliability in noise-only conditions.

5.3 Results – Quality of Descriptions in Overlapping Conditions

This subsection presents the evaluation results, comparing the audio descriptions produced by the model EchoLLM with those generated by a two-stage pipeline (AST+LLM). This comparison is critical for practical deployment, as it tests whether systems can generate accurate, fluent, and semantically faithful descriptions that reflect genuine understanding of the audio scene. Both approaches were evaluated using two protocols as described in Section 5.1.

5.3.1 Results from LLM-as-a-Judge Evaluation. Table 3 compares the baseline AST+LLM pipeline with EchoLLM under the human evaluation protocol described in Section 5.1. For human activity accuracy, EchoLLM achieved an average score of 4.92, with 94.15% of clips rated at the maximum score of 5, compared to 4.69 and 86.72% for AST+LLM. For background accuracy, EchoLLM obtained an average score of 4.64, with 89.25% of clips rated at 5, outperforming the baseline at 4.50 and 83.93%. The largest improvement was observed in overall semantic expression quality, where EchoLLM reached an average of 4.72 and 85.45% rated at 5, substantially higher than AST+LLM at 4.22 and 64.89%. These results demonstrate that EchoLLM improves accuracy for both activity and background recognition while also generating descriptions that are more coherent, contextually aligned, and semantically faithful to the reference.

It is also worth noting that many predictions scored as 4 by ChatGPT-4o were, upon manual inspection, semantically equivalent to the ground truth and could reasonably have been rated as 5. This indicates that the reported proportions of maximum scores represent a conservative estimate of system performance, reflecting the evaluator's tendency to under-score borderline cases. In summary, EchoLLM consistently surpasses the AST+LLM pipeline across all evaluation dimensions. While classifier+LLM can match event-level accuracy, it reduces audio to discrete labels that limit contextual detail, often introduces hallucinated content during text generation, and yields lower semantic fidelity in the resulting narratives. EchoLLM model delivers higher recognition accuracy and superior sentence-level expression, underscoring the value of generation for practical deployment where natural, high-quality descriptions of domestic audio scenes are required.

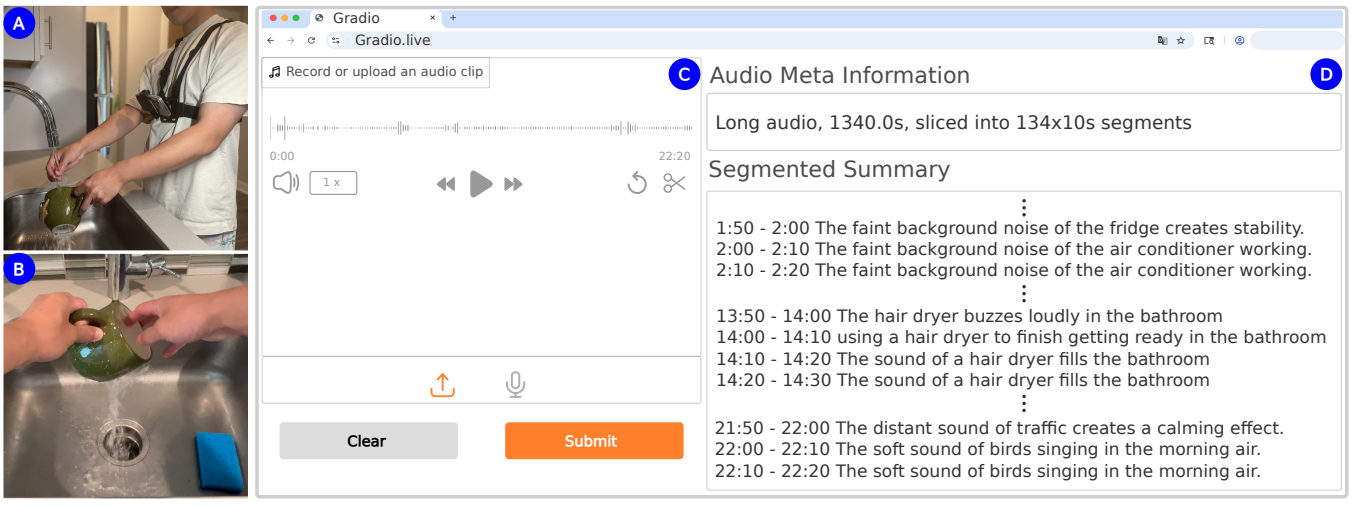

Figure 4: Example of the data collection and real-time inference interface. (A–B) Participants performed daily activities while carrying a chest-mounted smartphone that continuously recorded video. (C) The web-based interface supports real-time audio recording as well as uploading of audio recorded by participants. (D) Uploaded audio is automatically segmented into 10-second intervals, and the system generates corresponding moment-level descriptions.

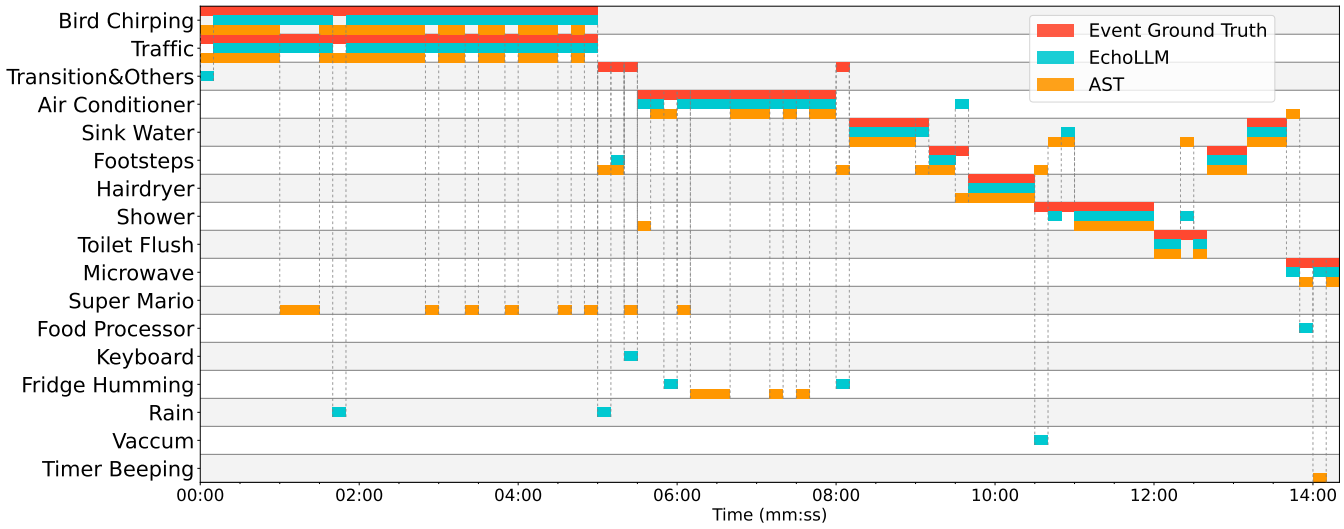

Figure 5: Timeline visualization of event ground truth, EchoLLM predictions, and AST predictions. EchoLLM demonstrates closer temporal alignment with annotated events compared to AST.

5.3.2 Results from Semantic Similarity Metrics. In this evaluation, each predicted sentence was compared with its corresponding ground-truth description to measure semantic similarity. EchoLLM achieved the highest performance on both metrics, with a BERTScore-F1 of 0.9075 and a Sentence-BERT similarity of 0.7030, compared to 0.8846 and 0.5942 for AST+LLM. These gains indicate that EchoLLM produces descriptions that are not only closer in meaning to the reference but also more coherent at the sentence level. For lifelogging applications, this capability is critical: reliable summaries must preserve the semantics of daily activities and their background context in natural language form, enabling outputs that are both accurate for automated analysis and understandable to end users.

5.4 Real-world Audio Scene Validation

Although EchoLLM was trained and evaluated on a synthesized dataset, its applicability must be verified in realistic, unconstrained settings. This validation examines the model's ability to recognize and narrate human activities from in-home recordings collected under natural environmental conditions. Importantly, none of these recordings were part of the training data, allowing the evaluation to test generalization across new users, diverse devices, and varied home environments—factors that are essential for deploying lifelogging systems in practice.

5.4.1 Data Collection and Setup. Three participants were recruited and asked to perform a variety of common activities in their own homes. Each participant carried a smartphone at chest height, as illustrated in Figure 4, which continuously recorded audio for approximately 15–20 minutes. A web-based interface was developed, as illustrated in Figure 4, to enable convenient real-world testing. The interface can be accessed from any device with a microphone over Wi-Fi, allowing participants to use their personal smartphones or other devices. Once recording is completed, users can submit the audio through the interface, which directly calls the trained EchoLLM model and returns inference results in real time. This design provides a lightweight and accessible means of deploying the system across diverse devices and environments without requiring specialized hardware.

The recordings were segmented into 10-second intervals and annotated by the researchers, who provided a brief descriptive phrase for each segment to capture the participant's activity. Both EchoLLM and the AST baseline processed the same recordings and produced a predicted activity label for each segment. Since the AST baseline model already can output discrete activity labels, its predictions were compared directly to the annotated ground truth. For EchoLLM, which can generate natural-language narratives, evaluation was restricted to the correctness of recognized activity labels to ensure a fair comparison. Accuracy, defined as the proportion of predicted labels matching the ground truth, was used as the evaluation metric.

5.4.2 Results. Across all participants, the system achieved an average activity recognition accuracy of 88.53%, with individual accuracies ranging from 84.1% to 92.0%. Recognition was generally reliable for longer-duration and acoustically distinctive activities (e.g., vacuuming, dishwashing), whereas short or low-intensity activities were more prone to substitution or omission. An example from one participant is shown in Figure 5 for visualization. These results represent raw performance obtained without any post-processing or filtering. The purpose of this evaluation is to demonstrate the system's ability to generalize in practical settings, using participants' own smartphones and varied home environments. The goal here is not to optimize accuracy but to establish feasibility under realistic conditions, with performance examined at each stage of the pipeline. Substantial improvements are expected once filtering strategies, such as those described in Section 3.2, are applied. Overall, these findings demonstrate that EchoLLM can be robustly deployed in real-world scenarios, providing a foundation for lifelogging applications.

Metric	EchoScriptor	Method 1 AST+LLM+GPT	Method 2 AST+LLM+GPT	Method 3 AST+LLM+GPT	E. vs M.2 Δ (Abs.)	E. vs M.2 %Δ
Classification accuracy						
Precision	**0.92**	0.71	0.85	0.90	+0.07	+8.2
Recall	**0.94**	0.49	0.74	0.68	+0.20	+27.0
F1-score	**0.92**	0.62	0.79	0.76	+0.13	+16.5
Semantic similarity						
BERTScore	**0.88**	0.87	0.86	0.87	+0.02	+2.3
Sentence-BERT	**0.68**	0.56	0.55	0.59	+0.13	+23.6
Alignment						
Activity Alignment	**0.81**	0.46	0.65	0.59	+0.16	+24.6

Table 4: Summary-level evaluation comparing EchoScriptor with three AST-based baselines. Method 1, Method 2, and Method 3 correspond to the pipelines illustrated in Figure 6. Absolute and percentage improvements are computed as Δ = EchoScriptor − Method 2 and $\%\Delta = 100 \times \frac{\text{EchoScriptor} - \text{Method 2}}{\text{Method 2}}$. Bold values indicate the best score in each row. E. represents EchoScriptor, and M.2 represents Method 2.

6 Narrative Summaries Quantitative Evaluation

While previous sections focused on controlled datasets and moment-level metrics, it is equally important to evaluate EchoScriptor at the level of narrative summaries. This section describes the user study for data collection and quantitatively evaluates the accuracy and coherence of system-generated summaries by comparing them with the AST+LLM+GPT baseline and human-annotated ground truth. The evaluation examines whether EchoScriptor can reliably capture both the content and temporal structure of daily activities, producing outputs suitable for lifelogging applications.

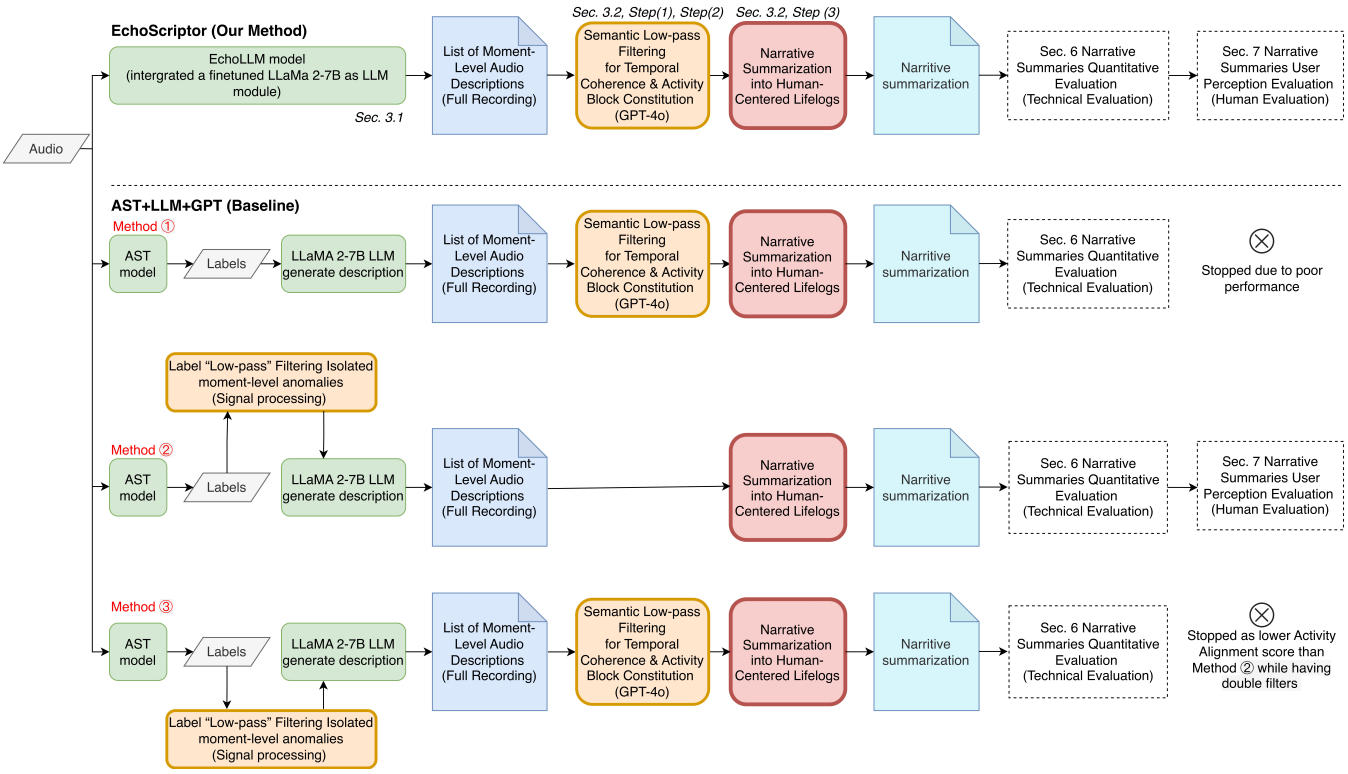

Figure 6: Pipeline comparison between EchoScriptor and the three AST-based baseline methods. Method 1: AST + LLaMa + Semantic Filter (GPT-4o) + Narrative Summarization (GPT-4o). Method 2: AST + Low-pass Filter + LLaMa + Narrative Summarization (GPT-4o). Method 3: AST + Low-pass Filter + LLaMa + Semantic Filter (GPT-4o) + Narrative Summarization (GPT-4o).

6.1 User Study Design for Narrative Summaries Quantitative Evaluation

To evaluate summary generation for lifelogging in real-world scenarios, a user study was conducted with ten participants. The study protocol was reviewed and approved by the University of Michigan Institutional Review Board (IRB), and all participants provided informed consent. The study was designed to capture realistic household activities across diverse environments. Participants carried out routine activities in their own homes, ensuring naturalistic behavior. Data were collected across ten distinct households with varied room layouts, appliance setups, and device configurations. Each home contributed one continuous session of 10–20 minutes, recorded using a chest-mounted iPhone 13 Pro that captured egocentric video with synchronized audio. Recordings took place at different times of day to increase diversity in acoustic conditions and activities. Only the audio stream was provided as input to EchoScriptor; the video stream was retained solely to support participant recall and visualization during the later user perception study. Participants viewed the video simply to recall and verify the activities as a reference they had performed. This design choice reflects the inherent challenge of inferring activities from audio alone, since many actions produce weak or ambiguous acoustic cues. After each recording, participants were asked to write a brief paragraph describing the activities performed and relevant contextual details. Some participants initially provided only keywords or short bullet points, in which case a follow-up prompt requested a more complete paragraph. These human-authored descriptions served as a comparison condition in the later evaluation phase.

6.2 Summaries Quantitative Evaluation Results

From the user study, ten videos were collected, each associated with three summaries: one generated by EchoScriptor, one by the AST+LLM+GPT baseline, and one authored by the participant. The goal of this evaluation is to quantitatively assess the accuracy of system-generated summaries at the summary level, focusing on whether they capture the activities and background events present in the recordings. Human-authored summaries and annotated egocentric videos served as references. Ground-truth event lists were derived from these annotations using short descriptive phrases. For both EchoScriptor and AST+LLM+GPT, comprehensive activity lists were manually extracted from their generated summaries to enable direct comparison with the ground truth.

The predicted and ground-truth audio events lists were compared using three categories of metrics: (i) standard classification measures (precision, recall, F1-score), which evaluate activity overlap without considering order; (ii) semantic similarity measures (BERTScore and Sentence-BERT, described in Section 5.3.2); and (iii) a sequence alignment metric. The alignment score was computed using the normalized edit distance between the predicted and ground-truth activity sequences, defined as

$$\text{Align}(P, G) = 1 - \frac{\text{EditDist}(P, G)}{\max(|P|, |G|)}, \tag{2}$$

where $\text{EditDist}(P, G)$ is the Levenshtein edit distance and $|\cdot|$ denotes sequence length. Unlike precision, recall, and F1-score, which ignore temporal order, the alignment score explicitly captures both correct identification of activities and preservation of their temporal sequence relative to the reference annotations.

To better illustrate the comparison, Figure 6 summarizes the detailed pipeline used for EchoScriptor and the AST+LLM+GPT baselines. EchoScriptor first produces moment-level audio descriptions using the EchoLLM model. These descriptions are then processed by the semantic "low-pass" filtering (Section 3.2 Step (1)) and activity-block construction modules (Section 3.2 Step (2)). These processes jointly remove short, noisy, or incoherent detections and enforce temporal consistency. Finally, the filtered moment-level descriptions are passed to the GPT-4o-based narrative summarization module (Section 3.2 Step (3)) to generate the final human-centered activity summaries. For the AST+LLM+GPT baseline, the first configuration, Method 1, replicates the full EchoScriptor pipeline: AST-predicted labels are converted into moment-level descriptions and then processed through the same semantic filtering and narrative-summarization stages. This configuration yielded limited performance, as shown in Table 4. In order to strengthen the baseline, two additional variants were further examined. Because the AST model produces a categorical label stream, the smoothing operations in the Section 3.2 Step (1 & 2) can be applied more effectively using a conventional label-level low-pass filter. Method 2 applies this smoothing before generating moment-level descriptions with LLaMA-2-7B. Method 3 extends this approach by adding EchoScriptor's semantic filter on top of the label-level filter, resulting in a two-stage filtering pipeline.

All three variants were evaluated under identical metrics, with results summarized in Table 4. Methods 2 and 3 achieved similar overall performance, although Method 3 showed slightly reduced activity-alignment accuracy. Consequently, Method 2 was selected as the baseline for the user-perception study in Section 7. It is possible that applying two filtering stages will remove information that should have been retained, resulting in unintended loss of relevant content. Notably, all methods, including every baseline variant, used the same GPT-4o narrative-summarization prompt, ensuring that differences at the summary level reflect upstream audio-understanding capabilities rather than variation in the summarization stage. Finally, the results in Table 4 show that EchoScriptor consistently outperforms the best AST+LLM+GPT baseline across all evaluation dimensions. In classification accuracy, EchoScriptor achieves precision of 0.92, recall of 0.94, and F1-score of 0.92, reflecting not only clear improvements over the baseline but also strong absolute performance. Comparable gains are observed in semantic similarity (BERTScore 0.88, Sentence-BERT 0.68) and temporal alignment (0.81), indicating that the system produces summaries that are both semantically faithful and chronologically consistent with the ground truth. Relative to the best AST+LLM+GPT baseline, EchoScriptor improves recall by +27.0%, F1-score by +16.5%, and alignment by +24.6%, demonstrating substantial advantages beyond incremental gains. Taken together, these results show that EchoScriptor is capable of generating high-quality narrative summaries that reliably capture daily activities and background context.

In the user study, participants were instructed to behave as they normally would at home and were not asked to perform any list of specific scripted actions. Some short, incidental transition sounds, such as opening doors or cabinets, or tearing packaging, did occur naturally but were not present in the EchoScriptor training dataset.

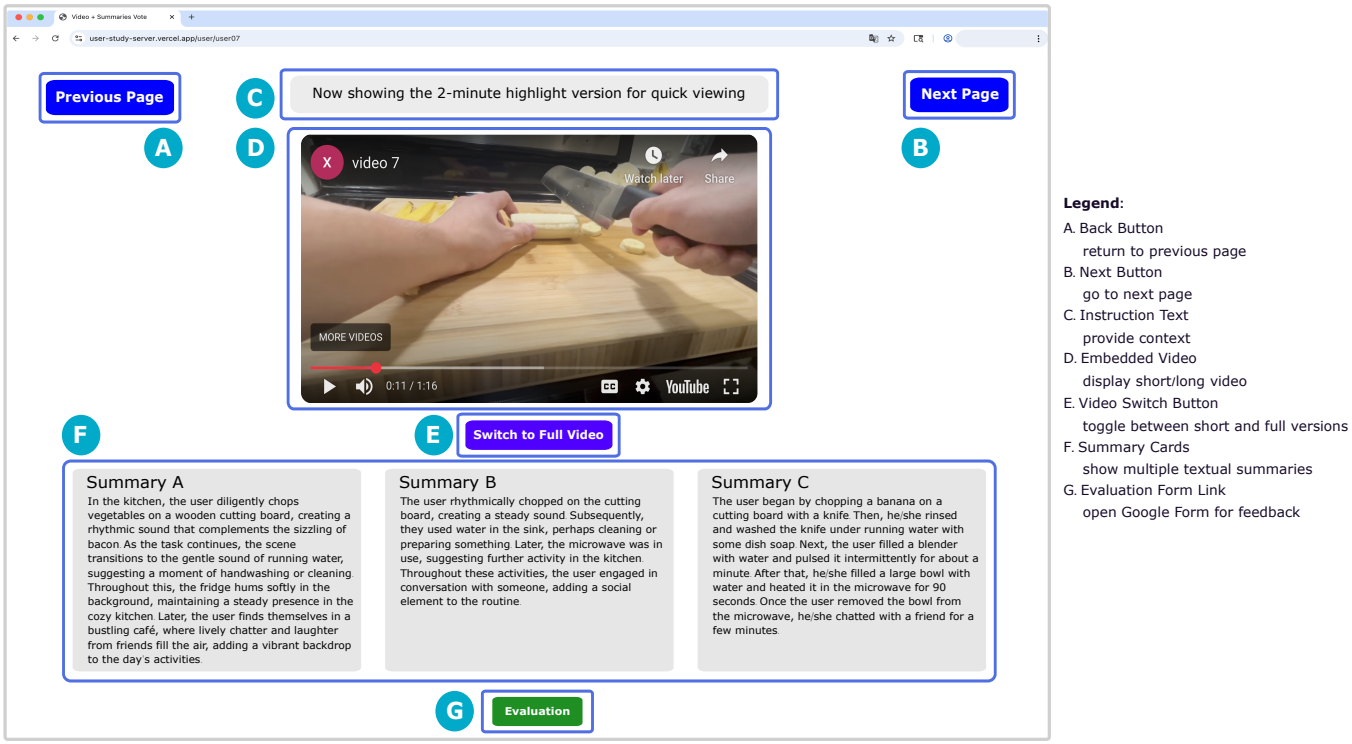

Figure 7: Custom web interface used in the user perception study. The interface includes a video player (highlight or full-length), anonymized summary cards, as well as navigation and evaluation controls.

Crucially, the system is designed to handle such occurrences. As described in the Semantic Filtering module (Section 3.2), the pipeline enforces temporal coherence, meaning that brief, isolated acoustic events are treated as anomalies and filtered out during the narrative construction. Therefore, while these sounds might appear in raw moment-level captions, they do not disrupt the detection of primary activities or the coherence of the final narrative summary. Longer or more sustained out-of-domain sounds, however, may be misinterpreted under the current model, which is an expected limitation given that the system does not achieve perfect accuracy. Addressing such cases is an area for future refinement. Continued improvements to the audio encoder and filtering strategies should expand the system's robustness to a wider range of acoustic conditions.

7 Narrative Summaries User Perception Evaluation

Although Section 5 evaluated moment-level performance and Section 6 examined narrative summaries with quantitative metrics, these analyses establish only the technical foundation. Equally important are human-centered qualities such as readability, coherence, trustworthiness, and perceived value in use. Since the intended application of EchoScriptor is lifelogging in a user-facing context, user experience is essential. To address this dimension, a qualitative evaluation was conducted in which participants reviewed the summaries for each video and compared EchoScriptor with a human-authored description and the AST+LLM+GPT baseline.

7.1 User Study Design

For each of the ten videos described in Section 6, three summaries were prepared for evaluation: a human written description, an EchoScriptor-generated summary, and an AST+LLM+GPT baseline summary. The user perception study involved twenty participants and was approved by the Institutional Review Board (IRB), with informed consent obtained from all participants. Each participant reviewed the corresponding video and audio recording from the study and then evaluated the three summaries through a questionnaire followed by a short interview.

A custom web application was developed to present each video alongside its three summaries, as shown in Figure 7. Participants accessed the interface remotely through a secure link and completed the task on their personal computers, with each session moderated via Zoom by a researcher. At the top of the interface, a video player displayed a two-minute highlight summarizing the main events with repetitive segments removed. A control label allowed participants to switch to the full-length recording at any time, with a status message indicating which version was active. The player supported standard functions, including pause, seek, and replay. Below the player, three anonymized summaries were displayed as separate cards labeled with A, B, and C. Label assignment and left-to-right order were randomized for each participant and each video to mitigate order effects. All summaries were blinded for authorship, and participants were not informed of the system that generated each one. To minimize stylistic bias, all texts were written in third-person narration using "the user" as the subject.

Table 5: Questionnaire for evaluating summary quality and perceived value in lifelogging contexts. Q1–Q6 and Q10 used a 7-point Likert scale (1 = strongly disagree, 4 = neutral, 7 = strongly agree). Q7–Q9 were rank-order items (1 = best, 3 = worst). Q11 was an open-ended free-text response.

ID	Aspect	Question
Overall Summary Quality (1–7 Likert-rated)		
Q1	Accuracy	The summary accurately described the activities performed during the session.
Q2	Precision	The summary is concise and avoids redundant or false information.
Q3	Clarity / Naturalness / Detail	The summary was easy to read, coherent, and included sufficient detail to understand what happened.
Q4	Chronological Order	The summary described the activities in the order they occurred.
Perceived Quality in Lifelogging (1–7 Likert-rated)		
Q5	Trustworthiness	I would feel confident relying on this summary as an accurate and trustworthy record of the event.
Q6	Perceived Utility	If used in a lifelogging or memory-support diary, this summary would help me recall events from my daily life.
Rank-order items (1 = best, 3 = worst)		
Q7	Clarity	Rank the summaries by how clearly they helped you understand the session.
Q8	Helpfulness	Rank the summaries by how helpful they would be for supporting recall in a lifelogging diary.
Q9	Comfort	Rank the summaries by how comfortable you would feel revisiting them later to recall your past activities.
General evaluation items		
Q10	Perceived Value	If a lifelogging app could automatically record your activities for you, eliminating the need to write them down yourself, how valuable would you find this in making it easier to keep an accurate record of your day?
Q11	Open-ended	Please briefly explain why you preferred the summary you chose.

Human-authored descriptions originally written in the first person were converted into this style without altering their content. These measures ensured that participant evaluations reflected the quality of the summaries rather than authorship or stylistic cues.

7.1.1 Participants and Experimental Design. Twenty participants from a university community took part in this phase (8 female, 12 male; $M = 24.95$, $SD = 3.67$). All were fluent or professionally proficient in English. To avoid self-evaluation, participants who had contributed to the video collection did not assess their own recordings. This design ensured that all evaluations were independent and unbiased by participants' prior familiarity with their own activities. Each participant evaluated four of the ten videos, with assignments randomized and balanced so that each video received eight independent evaluations. For each assigned video, participants rated all three summaries, yielding 80 video-by-participant trials and 240 summary ratings in total. Each session lasted approximately 40 minutes, including onboarding, rating tasks, and debriefing. This allocation yields eight independent ratings per video for each summary type. It is consistent with common practice for participant sample sizes in Human-Computer Interaction research [10], aligns with summarization benchmarks that collect eight judgments per summary [25], and supports crossed participant-by-item analysis using linear mixed-effects models, which are recommended for this type of design [9, 92].

7.1.2 Questionnaire. After viewing the assigned video and reading all three summaries, participants answered eleven questions grouped into three categories (Table 5). Q1–Q4, rated on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree), assessed overall summary quality in terms of accuracy, precision, clarity, and chronological order. Q5–Q6, also Likert-rated, evaluated perceived value for lifelogging, specifically trustworthiness and memory support. Q7–Q9 were rank-order items (1 = best, 3 = worst), comparing the three summaries on clarity, helpfulness for recall, and comfort when revisiting. Finally, Q10 measured the perceived value of automatic summarization for lifelogging, and Q11 was an open-ended prompt for free-text comments. To reduce bias, participants were instructed not to let minor factual errors disproportionately influence ratings of unrelated qualities such as coherence or readability, and to evaluate each item according to its specific focus. All questions were presented in statement form, with higher ratings indicating stronger agreement and corresponding to better performance.

7.2 Quantitative Results of User Perception

The analysis aimed to evaluate whether the three types of summaries (Human, AST+LLM+GPT baseline, and EchoScriptor) differed in perceived quality. For Q1–Q6 (7-point Likert ratings), linear mixed-effects models (LMMs) were applied to account for the repeated-measures design, in which each participant rated four videos and each video was evaluated by eight participants. Unlike standard ANOVA, which assumes independence and sphericity, LMMs explicitly account for dependencies by modeling participant and video as random intercepts, with summary type as the fixed effect of interest. This approach yields robust significance tests under the crossed design. The analysis tested whether average ratings differed by summary type while controlling for individual differences in rating tendencies and for variation in the baseline difficulty of the videos. 95% confidence intervals (CIs) rather than standard deviations (SDs) are reported, as CIs reflect the uncertainty of the estimated means rather than the variability of individual responses. This choice emphasizes the precision of mean differences across summaries and aligns with inferential analyses, making it clearer whether the observed differences are statistically meaningful.

The consistency of participant ratings was also examined. Each video–summary cell (e.g., the Human summary for Video 1) was evaluated by eight participants, introducing variability both across raters and across videos. Simple averages may therefore obscure the degree of agreement among raters. To quantify this agreement, the intraclass correlation coefficient ICC(1,k) was computed, a one-way random-effects measure of absolute agreement among multiple raters nested within targets. Reliability in this context reflects the

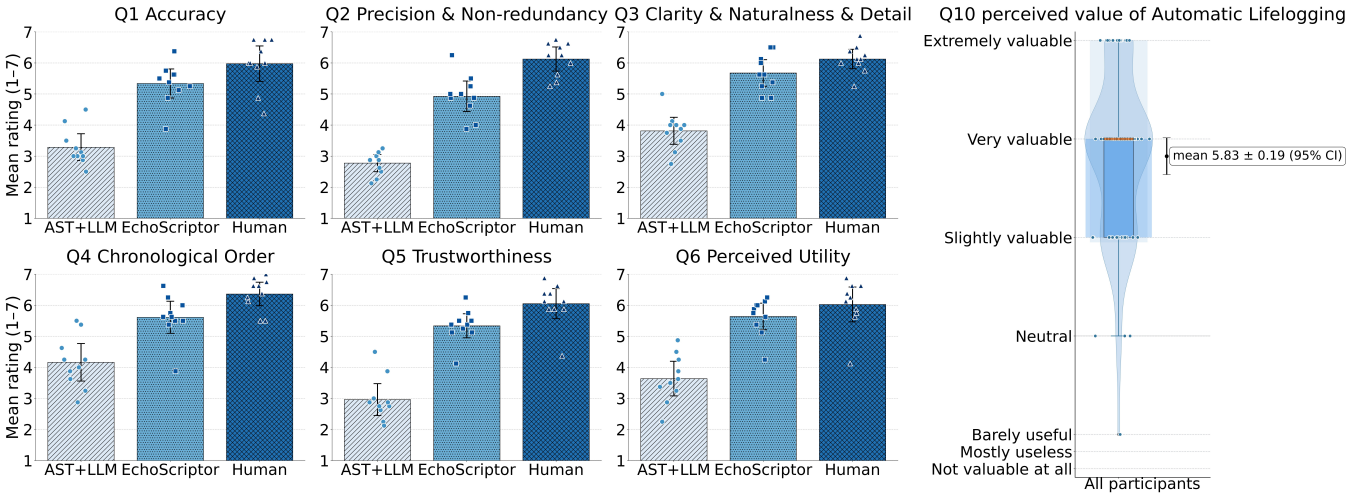

Figure 8: Mean participant ratings (Q1–Q6) of summary quality across three conditions: AST+LLM+GPT baseline, EchoScriptor, and human-written summaries. Each bar shows the average rating on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). Error bars represent 95% confidence intervals of the mean, meaning they show the uncertainty in the average rating across videos. Individual points mark the mean rating for each video, meaning they display how results varied from video to video rather than just the overall average. The Q10 panel shows a violin plot of participant ratings for the perceived value of automatic lifelogging, where the width of the violin indicates the distribution density, the box spans from the first quartile (Q1, 25th percentile) to the third quartile (Q3, 75th percentile), the horizontal orange line marks the median, and individual dots represent participant responses.

consistency of scores assigned to the same summary: high ICC values (close to 1) indicate strong agreement, whereas low or negative values indicate disagreement and less reliable mean ratings. ICCs were calculated separately for each question and summary type using $k = 8$ raters per cell, with 95% confidence intervals estimated via bootstrapping across videos.

Moreover, to evaluate participants' direct preferences among the three systems, participants ranked the Human, EchoScriptor, and AST+LLM+GPT summaries on clarity (Q7), helpfulness (Q8), and comfort (Q9). Pairwise preferences were analyzed by comparing, for each participant, how often one system was ranked higher than another (ties were noted and excluded from Wilcoxon tests). Wilcoxon signed-rank tests with Holm adjustments were used to assess the significance of pairwise differences. In addition, Friedman tests were applied across all three systems per question, treating participants as the blocking factor, to test for overall effects. All participants completed the assigned ratings, and no data were excluded from these analyses.

7.2.1 Summary Quality Results (Q1–Q6). Figure 8 presents participant ratings across Q1–Q6. Human-written summaries consistently received the highest ratings, with per-question means ranging from 5.98 to 6.36 on the 7-point scale. EchoScriptor ranked in the mid-range, with means from 5.34 to 5.68, while the AST+LLM+GPT baseline received the lowest ratings, with means from 2.78 to 4.16. For example, on Q1 (Accuracy), human summaries averaged 5.98 (95% CI [5.72, 6.23]), EchoScriptor averaged 5.34 (95% CI [5.02, 5.65]), and the AST+LLM+GPT baseline averaged 3.29 (95% CI [2.93, 3.64]). A similar pattern was observed for Q2–Q6. When collapsing across all six evaluation questions, the overall averages reinforced this pattern: human summaries received the highest ratings (mean = 6.11, 95% CI [6.01, 6.21]), followed by EchoScriptor (mean = 5.42, 95% CI [5.30, 5.54]), and the AST+LLM+GPT baseline (mean = 3.44, 95% CI [3.28, 3.60]).

Pairwise contrasts confirmed this pattern as shown in Table 6: human summaries were rated significantly higher than both EchoScriptor and AST+LLM+GPT on every dimension, and EchoScriptor was rated significantly higher than AST+LLM+GPT (all $p_{\text{adj}} < .05$). The only exception was Q6 (Perceived Utility), where ratings for human (6.03, 95% CI [5.65, 6.40]) and EchoScriptor (5.64, 95% CI [5.26, 6.02]) did not differ significantly ($p = .062$). Overall, these results demonstrate that EchoScriptor substantially improves over the AST+LLM+GPT baseline but still falls short of human performance, except in Q6-perceived-utility, where it approaches parity. Across all six quality dimensions (Q1–Q6), EchoScriptor received consistently positive ratings. These results indicate that participants found EchoScriptor's outputs not only technically sound but also sufficiently clear, trustworthy, and useful to be acceptable in practice. Rather than being dismissed as an inferior baseline, EchoScriptor was evaluated as a viable, high-quality alternative that participants could imagine relying on in real use.

7.2.2 Inter-Rater Reliability of Ratings. On average across Q1–Q6, Human summaries achieved the highest reliability (ICC = 0.67, range 0.11–0.83), indicating that participants generally agreed in their judgments of higher-quality summaries. Reliability was lowest for clarity (Q3, ICC = 0.11), suggesting greater divergence in perceptions of clarity. EchoScriptor showed moderate reliability overall (ICC = 0.47, range 0.28–0.59), indicating that participants were somewhat consistent when evaluating its outputs. By contrast, the AST+LLM+GPT baseline exhibited very low reliability (ICC = 0.06, range −0.99–0.50), with several questions yielding

Table 6: Linear mixed-effects model results for Q1–Q6. All models showed a significant main effect of summary type (LR $\chi^2[2] = 91–197$, all $p < .001$). Positive estimates indicate the first method in the contrast received higher ratings.

Q1 Accuracy

Contrast	Estimate	z	p_{adj}
Human – EchoScriptor	0.64	3.02	.003
Human – AST+LLM+GPT	2.69	12.71	< .001
EchoScriptor – AST+LLM+GPT	2.05	9.70	< .001
Human ≫ EchoScriptor ≫ AST+LLM+GPT			

Q2 Precision

Contrast	Estimate	z	p_{adj}
Human – EchoScriptor	1.20	6.36	< .001
Human – AST+LLM+GPT	3.35	17.75	< .001
EchoScriptor – AST+LLM+GPT	2.15	11.39	< .001
Human ≫ EchoScriptor ≫ AST+LLM+GPT			

Q3 Clarity

Contrast	Estimate	z	p_{adj}
Human – EchoScriptor	0.45	2.27	.023
Human – AST+LLM+GPT	2.31	11.69	< .001
EchoScriptor – AST+LLM+GPT	1.86	9.41	< .001
Human ≫ EchoScriptor ≫ AST+LLM+GPT			

Q4 Order

Contrast	Estimate	z	p_{adj}
Human – EchoScriptor	0.75	3.61	< .001
Human – AST+LLM+GPT	2.20	10.60	< .001
EchoScriptor – AST+LLM+GPT	1.45	6.99	< .001
Human ≫ EchoScriptor ≫ AST+LLM+GPT			

Q5 Trustworthiness

Contrast	Estimate	z	p_{adj}
Human – EchoScriptor	0.71	3.63	< .001
Human – AST+LLM+GPT	3.09	15.75	< .001
EchoScriptor – AST+LLM+GPT	2.38	12.11	< .001
Human ≫ EchoScriptor ≫ AST+LLM+GPT			

Q6 Perceived Utility

Contrast	Estimate	z	p_{adj}
Human – EchoScriptor	0.39	1.87	.062
Human – AST+LLM+GPT	2.39	11.52	< .001
EchoScriptor – AST+LLM+GPT	2.00	9.65	< .001
Human ≳ EchoScriptor ≫ AST+LLM+GPT			

near-zero or negative ICCs (e.g., Q2 Precision, ICC = −0.99), reflecting substantial disagreement when judging the lowest-quality summaries. Overall, these results indicate that participant ratings were most consistent for Human summaries, moderately consistent for EchoScriptor, and least consistent for AST+LLM+GPT outputs.

7.2.3 Pairwise Preference Rankings (Q7–Q9). Table 7 summarizes participants' pairwise preferences when ranking the three systems on clarity (Q7), helpfulness (Q8), and comfort (Q9). The mean rank order was Human (1.45) ≫ EchoScriptor (1.72) ≫ AST+LLM+GPT (2.83). For Q7 (Clarity), Human was preferred over EchoScriptor (79% of participants, p=0.017) and strongly over AST+LLM+GPT (p<.001). At the same time, EchoScriptor also significantly outperformed AST+LLM+GPT (p<.001). A similar pattern was observed for Q8 (Helpfulness). For Q9 (Comfort), participants again favored Human over AST+LLM+GPT (p<.001) and EchoScriptor over AST+LLM+GPT (p<.001), with no reliable difference between Human and EchoScriptor. Friedman tests confirmed strong overall effects for each question (all p<.001). When pooling across Q7–Q9, Human summaries were significantly preferred over EchoScriptor, and EchoScriptor was strongly favored over AST+LLM+GPT. Taken together, these rankings suggest that while Human summaries remain the best, EchoScriptor is perceived as much closer to Human quality and substantially better than the baseline. Participants often placed EchoScriptor on par with Human summaries for subjective dimensions such as helpfulness and comfort. This indicates that EchoScriptor's outputs are not only technically stronger than the baseline but also socially acceptable and practically usable.

7.2.4 Perceived Value of EchoScriptor. As shown in Figure 8, the Q10 results suggest that participants generally perceive automatic lifelogging as highly valuable, with most responses clustered at the upper end of the scale. The high median and the narrow confidence interval (5.83 ± 0.19) indicate consistent agreement, highlighting a broad understanding among participants that such a system could meaningfully support daily life. Only a few neutral or negative responses appeared, reinforcing the overall positive view.

7.3 Qualitative Findings from Open-Ended Question

Participants' open-ended reflections revealed four interconnected priorities that extend beyond numerical ratings and point to what users expect from lifelogging summaries in practice. First, accuracy was consistently identified as the most critical requirement. Users stressed that summaries must faithfully represent recorded events, as even minor factual errors undermine trust and compromise their value as external memory aids, particularly in contexts such as forgetfulness or cognitive decline. One participant noted: *"I prefer Summary B. Although it doesn't provide much detail on the activities, it is the most accurate description of the activities that took place in the video. The other summaries include some embellished details that provide a more vivid description but are not necessarily true."*

Second, participants valued richness and readability. They preferred detailed accounts that captured both activities and surrounding context more fully than sparse human-written notes, yet they also emphasized that richness must be delivered with clarity, avoiding redundancy or excessive length that would make summaries difficult to skim. For example, baseline expression as *"The user moves with purpose, the sounds of their culinary activity blending with the ambient hum of kitchen appliances, crafting a vibrant and lively atmosphere. The day unfolds with a seamless blend of domestic tasks, each sound painting a vivid picture of daily life."* vs EchoScriptor expression *"Later, they turned their attention to the kitchen, where the steady rhythm of chopping on a cutting board indicated preparation for a meal."* Another example highlighted this contrast: the baseline produced descriptions such as *"The day begins with the soothing sound of water flowing steadily from the kitchen sink, creating a calm atmosphere as the user engages in domestic tasks,"* whereas EchoScriptor generated a more direct, factual account such as *"In the bathroom, the user engaged in a grooming routine, using an electric shaver, razor, and toothbrush."* Interestingly, one participant expressed a preference for slightly more poetic and dramatic descriptions from the baseline. For example: *"The user ascends the*

Table 7: Pairwise ranking analysis for Q7–Q9 (1 = best, 3 = worst). Each row shows the number of participants whose mean ranks favored system A over B ("Wins A" vs "Wins B"), expressed as counts and percentages, with *Ties* indicating participants who gave equal mean ranks to both systems (and who are excluded from Wilcoxon tests). Holm-adjusted p values come from Wilcoxon signed-rank tests on per-participant mean ranks ($n = 20$). Lower ranks indicate better performance. Stars: * $p<.05$, ** $p<.01$, *** $p<.001$. Friedman tests (shaded rows) treat participant as the blocking factor and provide overall comparisons. Conclusion symbols: ≫ = significantly better; ≳ = not significantly different.

Question	Pair	Wins A	%A	Wins B	%B	Ties	Holm-p	Conclusion
Q7 Clarity	Human–EchoScriptor	11	78.6	3	21.4	6	1.66×10^{-2}	Human > EchoScriptor *
	Human–AST+LLM+GPT	20	100.0	0	0.0	0	6.0×10^{-6}	Human > AST+LLM+GPT ***
	EchoScriptor–AST+LLM+GPT	20	100.0	0	0.0	0	4.0×10^{-6}	EchoScriptor > AST+LLM+GPT ***
					Friedman: χ^2=34.162, p=3.82×10^{-8}			⇒ Human ≫ EchoScriptor ≫ AST+LLM+GPT
Q8 Helpfulness	Human–EchoScriptor	12	75.0	4	25.0	4	5.07×10^{-2}	Human ≈ EchoScriptor
	Human–AST+LLM+GPT	18	90.0	2	10.0	0	1.09×10^{-4}	Human > AST+LLM+GPT ***
	EchoScriptor–AST+LLM+GPT	19	95.0	1	5.0	0	2.67×10^{-4}	EchoScriptor > AST+LLM+GPT ***
					Friedman: χ^2=24.105, p=5.83×10^{-6}			⇒ Human ≳ EchoScriptor ≫ AST+LLM+GPT
Q9 Comfort	Human–EchoScriptor	11	73.3	4	26.7	5	1.07×10^{-1}	Human ≈ EchoScriptor
	Human–AST+LLM+GPT	19	95.0	1	5.0	0	1.1×10^{-5}	Human > AST+LLM+GPT ***
	EchoScriptor–AST+LLM+GPT	19	100.0	0	0.0	1	1.1×10^{-5}	EchoScriptor > AST+LLM+GPT ***
					Friedman: χ^2=28.892, p=5.32×10^{-7}			⇒ Human ≳ EchoScriptor ≫ AST+LLM+GPT
Overall Q7–Q9	Human–EchoScriptor	12	66.7	6	33.3	2	3.85×10^{-2}	Human > EchoScriptor
Overall Q7–Q9	EchoScriptor–AST+LLM+GPT	19	100.0	0	0.0	1	8.0×10^{-6}	EchoScriptor > AST+LLM+GPT
								Overall conclusion: Human ≫ EchoScriptor ≫ AST+LLM+GPT

Overall mean rank (Q7–Q9 pooled; lower is better): EchoScriptor = 1.72, Human = 1.45, AST+LLM+GPT = 2.83.

stairs, their footsteps resonating on the wooden steps, leading them to a bustling kitchen."

Third, participants expressed concern about redundancy and hallucination in the baseline summaries (e.g., "joyfully vacuuming"), noting that overly elaborate language or invented details detracted from reliability. Conciseness, therefore, emerged as the default expectation, with more narrative or expressive phrasing acceptable only when it added meaningful context without compromising factual accuracy. For instance, one participant wrote: "After that, the user filled a large bowl with water and heated it in the microwave for 90 seconds. Once the user removed the bowl from the microwave, the user chatted with a friend for a few minutes." In contrast, the baseline produced a highly stylized but incorrect description, such as: "Throughout this, the fridge hums softly in the background, maintaining a steady presence in the cozy kitchen. Later, the user finds themselves in a bustling café, where lively chatter and laughter from friends fill the air, adding a vibrant backdrop to the day's activities," which misidentified the activity and context. EchoScriptor, however, generated a more grounded and contextually aligned summary: "Later, the microwave was in use, suggesting further activity in the kitchen. Throughout these activities, the user engaged in conversation with someone, adding a social element to the routine." Finally, personalization was regarded as essential. For example, noting the topic of a conversation ("talked about graduate school") or referencing affective states ("energetic morning", "it is fun") were consistently rated highly because they aligned with how participants naturally recorded their own experiences for a home diary. Participants highlighted that lifelogging should not only document actions (e.g., "walking around the room") but also reflect intentions, emotions, and conversational topics (e.g. "walking around the room while planning tomorrow's work"), which made the summaries feel closer to personal diary entries and therefore more engaging and useful. Together, these findings suggest that effective lifelogging systems must balance factual reliability, sufficient detail, clarity, and personalization to create records that are not only trustworthy but also meaningful to the individual.

Overall, the whole user study demonstrates that EchoScriptor delivers narrative summaries that are both technically strong and positively received by users. Quantitative results demonstrated that it consistently outperformed the AST+LLM+GPT baseline across accuracy, precision, clarity, and trustworthiness, and approached human performance on perceived utility. Qualitative feedback reinforced these findings, highlighting accuracy as the non-negotiable foundation, the value of detail when paired with readability, the expectation of conciseness without redundancy or hallucination, and the importance of personalization that captures context and affective states. These results show that EchoScriptor is not only a technical advance in audio-based narrative generation but also a viable, user-acceptable step toward future lifelogging applications.

8 Discussion

This section examines the system along these several dimensions: (1) the feasibility of real-time deployment; (2) its trustworthiness and privacy properties; (3) the comparative strengths of audio-only versus multimodal sensing; (4) the technical limitations that motivate future research; and (5) the practical applications and broader implications.

8.1 Interactive Website Frontend with Real-time Deployment

The real-time deployment was designed to demonstrate not only the technical feasibility of the proposed system but also its usability in everyday contexts. A web-based interactive front-end was developed for users (Figure 9). The interface is globally accessible, allowing users to interact with the system directly through

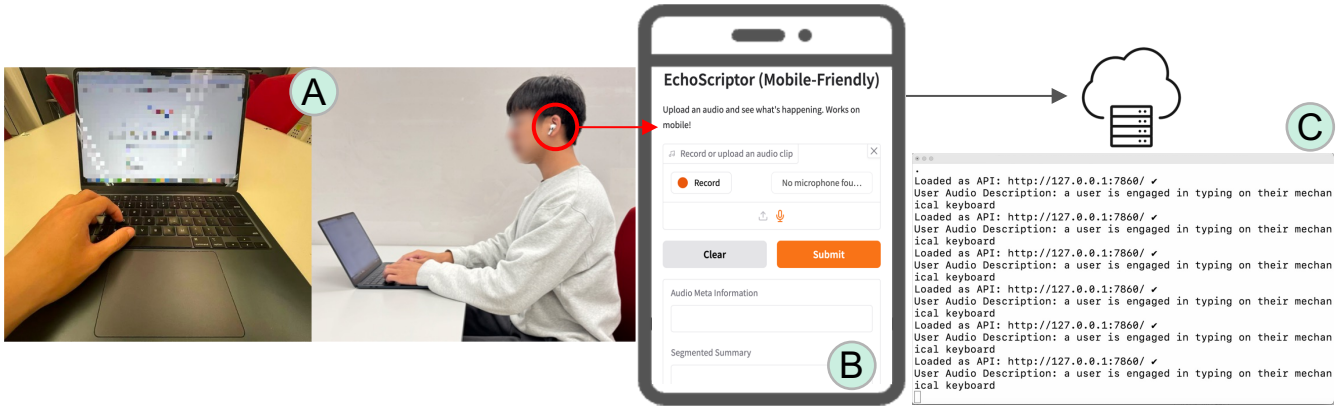

Figure 9: Real-time deployment of EchoScriptor. (A) A user types on a laptop while wearing earbuds as the recording device. (B) The mobile-friendly web interface allows audio capture or file upload and returns either streaming moment-level descriptions or session-level summaries. (C) Audio is processed on the server, with results delivered as natural-language activity descriptions.

any standard browser. Upon access, users can select from the microphone devices available to their browser to capture audio for analysis, such as those integrated into phones, headphones, smart speakers, or smartwatches. The system provides two modes of audio input. Users may either record audio in real time or upload existing recordings from local storage. Common file formats such as MP4, MOV, WAV, and other widely used audio and video types are supported. The system also offers two output modes. It can produce moment-level descriptions live as audio is captured, or process longer recordings, such as a 20-minute clip, in a single upload and return the full set of results once processing is complete. The goal of this deployment is to bridge algorithmic advances with real-world use, illustrating how lifelogging systems can be delivered in an accessible form that requires no specialized hardware or installation. By combining technical robustness with user-centered design, EchoScriptor provides a step toward practical, scalable applications of audio-based activity journaling in the future.

8.2 Trustworthiness and Privacy

Findings from our user study suggest that the long-term viability of continuous audio logging hinges on the balance between narrative utility and privacy preservation. Although EchoScriptor touches on this tension in the context of domestic lifelogging, the underlying challenges extend far beyond our current system and point to broader questions that the audio-based ubiquitous sensing, LLM-agent, and HCI communities may face. While participants generally perceived audio as less invasive than video, the continuous nature of sensing raises significant concerns regarding consent, bystander privacy, data security, and other issues discussed below. Widespread adoption of audio-driven narrative agents requires shifting focus from simple data capture to systems that prioritize privacy by design.

- Bystander Awareness and Privacy: A primary challenge for an always-on audio AI system is the "bystander problem", recording individuals who have not consented. Ethical deployment requires mechanisms that ensure fairness to those surrounding the user. Beyond physical indicators of sensing status (e.g., LEDs), future research must advance speaker disentanglement technologies. Rather than simply recording the environment, the system should ideally employ target-speaker extraction to isolate the user's voice while automatically suppressing or obfuscating non-enrolled speakers. This ensures that the resulting narrative reflects the user's perspective without infringing on the privacy of family members or the public.
- Risks in Embedding Representations: While high-level audio embeddings offer a degree of obfuscation compared to raw audio, the risks of model inversion must be addressed. Prior work demonstrates that embeddings can be reverse-engineered to recover speech content or paralinguistic markers (e.g., health status, emotional state). Future research should pursue privacy-preserving representation learning, developing embedding techniques retaining the semantic information necessary for summarization while mathematically stripping away biometric or identifiable features.
- Semantic Privacy Filtering: Current privacy mechanisms typically rely on coarse-grained binary controls (e.g., manual mute buttons), imposing a high cognitive burden on the user to anticipate sensitive moments. This reliance on manual intervention is inherently reactive, leaving privacy compromised before the user can act. To mitigate this, future architectures must advance toward context-aware semantic filtering to enhance user privacy. Privacy risks in domestic environments extend beyond intelligible speech; non-linguistic acoustic events, such as signs of medical distress, heated arguments, coughing,crying, or intimacy, carry high sensitivity even in the absence of words. Consequently, robust privacy frameworks must identify and suppress these acoustic categories at the sensor edge, ensuring such data is discarded prior to any downstream processing.
- Data Security and Protection: Reliance on centralized cloud architectures for continuous audio processing introduces inherent privacy risks. Transmitting raw audio data to servers expands the attack surface, increasing the chance of data

interception and unauthorized storage. While acceptable for research prototyping, widespread deployment requires strict security measures. Specifically, systems must enforce end-to-end encryption and transient processing protocols. Moreover, while the training of LLMs currently requires servers, the inference stage for user daily applications can be migrated to a smartphone or wearable device. This shift ensures that the data stays in the user's possession, replacing reliance on server policies with privacy by design.

8.3 Audio vs. Multi-modal Sensing

For the broader goal of comprehensive user activity understanding, audio sensing plays a pivotal role by capturing rich interaction details and environmental cues. Nevertheless, acoustic models face inherent limitations, particularly when detecting silent events or activities with ambiguous sonic signatures. To overcome these modality-specific constraints, future work can position EchoScriptor as a core acoustic reasoning module within a multimodal framework. Complementing audio-generated narratives with other privacy-preserving signals, such as lightweight IMU data, ambient device logs, or coarse location cues, would enable a more comprehensive interpretation of daily life. In another aspect, the choice between an audio-specific model and a multimodal foundation model reflects a balance of trade-offs. An audio-specialized or application-oriented model can be tuned to the characteristics of a specific task or dataset, enabling it to operate as an "expert" with strong performance and efficiency in that domain. In contrast, a multimodal model with iterative prompt or instruction tuning offers greater versatility and may generalize more effectively to unfamiliar scenarios. By retaining additional modalities, such models can provide complementary context when audio signals are weak, ambiguous, or unavailable, allowing modalities to support and correct one another. However, these benefits come with practical constraints. Multimodal models typically require greater computational resources, higher latency, and more demanding deployment conditions, which may limit their suitability for on-device or always-on use. In settings where edge devices do not have the capacity to run a full multimodal model, an audio-specific approach may remain the more appropriate solution. Future research needs to consider these trade-offs, such as performance, generalizability, resource demands, use cases, and user experience when determining the right balance between specialized and multimodal architectures.

8.4 Limitations and Future Work

While EchoScriptor establishes a promising technical foundation for audio-based lifelogging, it comes with several limitations that highlight opportunities for further improvements and extension.

A primary limitation lies in residual summarization errors. Addressing these errors requires audio encoders that capture finer acoustic detail and language models with stronger contextual reasoning. While EchoScriptor distinguishes activities from background, extending it to handle multiple overlapping events is an important direction to pursue. Finally, EchoScriptor's current reliance on fixed 10-second segments, imposed by the pretrained backbone, limits temporal continuity. Future models should directly process longer, variable-length audio streams for more coherent understanding.

Another limitation concerns acoustic diversity and the varieties of environmental conditions. To ensure robust generalization to real-world environments, the dataset is constructed with uncontrolled sources (e.g., YouTube, Freesound) rather than sterile studio Foley recordings. The dataset is "synthesized" in the sense that it is formed by overlapping sounds. As a result, the training samples inherently capture natural acoustic variations, including background noise, incidental sounds, and varying recording qualities, which contributed to the model's strong generalization to real homes in the user study. However, the current pipeline may not fully represent challenging acoustic complexities, such as heavy reverberation in large open spaces, multiple overlapping novel sound events, or far-field recordings of distant activities. While the current model generalizes well to typical domestic environments, our future work could improve robustness by incorporating impulse-response augmentation (e.g., via Pyroomacoustics) to simulate diverse room geometries or by expanding the training set to include wider-field spatial audio. Moreover, the dataset is synthesized and, despite manual verification, may still diverge from authentic human annotations to some extent. This limitation reflects a known challenge in the label synthesis and dataset construction field. We will need more advanced labeling algorithms or large-scale human-annotated datasets to achieve greater robustness and realism.

In addition, future work should examine the computational and energy efficiency of continuous audio processing to ensure that long-term deployment on wearable or edge devices is practical. Exploring lightweight model architectures and adaptive inference strategies may further reduce latency and power consumption while preserving summarization quality.

Finally, the current version of EchoScriptor is not intended or designed to be generalizable to all possible natural or unseen sounds. Although EchoScriptor incorporates pre-trained components, GAMA, CAV-MAE, and LLaMA, that possess broad open-world acoustic knowledge acquired from millions of diverse audio examples, its fine-tuning stage is performed on a domestic activity dataset. Importantly, this fine-tuning does not turn the system into a closed-set recognizer. Rather, its audio interpretation and narrative generation are adapted to household contexts, which are the intended use cases for lifelogging and memory-support applications. This domain-specific optimization prioritizes domestic accuracy because supporting users' daily recall was our primary design objective. Nevertheless, a systematic evaluation of zero-shot behavior on arbitrary non-domestic sounds remains an important avenue for future work. Future work should also explore how to balance such task-specific precision with the broader open-world recognition capabilities inherent in the underlying foundation models.

8.5 Practical Applications and Implications

The primary focus of this work is establishing a technical foundation for audio-based lifelogging, rather than the full design of end-user applications. Thus, our contribution lies in advancing the underlying modeling approach and establishing a systematic evaluation framework. Looking ahead, several practical applications and design directions emerge from this foundation. Lifelogging can support personal informatics by enabling users to reflect on daily routines, identify patterns in activity levels, or track behaviors that

evolve gradually over time. It can facilitate memory augmentation by helping individuals reconstruct the flow of their day, fill gaps in retrospective recall, and recover the context surrounding meaningful events that might otherwise be forgotten. In assistive scenarios, audio-based summaries can help individuals with memory impairments maintain awareness of what has occurred throughout the day without relying on intrusive sensing modalities. These daily summaries also provide caregivers or clinicians with a lightweight record of activities, offering a clearer picture of a patient's everyday functioning. Such contextual insight can improve diagnostic accuracy, support treatment planning, and allow providers to track how symptoms, such as changes in sleep, mobility, social engagement, or self-care, evolve over time. For older adults, this transparency can strengthen daily routines and enable more independent, dignified engagement in everyday life. These scenarios raise design considerations related to how summaries should be presented, how feedback should be tailored to different user needs, and how transparency and control should be integrated into the interface. Exploring these application-level questions in future deployments would help translate the technical contributions of EchoScriptor into usable, accessible, and ethically grounded real-world technologies.

9 Conclusion

We presented EchoScriptor, a system that converts raw household audio into coherent, human-readable narrative descriptions capturing both human activity and acoustic context. EchoScriptor comprises two main components: the EchoLLM large audio-language model, which generates moment-level audio descriptions, and the narrative construction pipeline, which produces longer session-level activity summaries. A large audio–description dataset was synthesized for training, and evaluations showed that EchoScriptor achieved 94.15% activity recognition and 89.25% background recognition, with 88.53% accuracy in real-world recordings. In a user study with 20 participants across 10 recordings, the generated summaries were rated as clearer, more trustworthy, and more useful than the baseline, approaching the perceived utility of human-written descriptions. A real-world deployment further demonstrated the system's accessibility and practicality. Together, these results position EchoScriptor as a promising step toward lifelogging and memory support technologies.

Acknowledgments

We sincerely thank all participants for their involvement in our user study. We also extend our gratitude to Sonal Kumar for their thoughtful collaboration, guidance, and practical support, which contributed meaningfully to the success of this research.

References

[1] Sharath Adavanne, Archontis Politis, Joonas Nikunen, and Tuomas Virtanen. 2018. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing* 13, 1 (2018), 34–48.

[2] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325* (2023).

[3] Mehmet Ali Arabacı, Fatih Özkan, Elif Surer, Peter Jančovič, and Alptekin Temizel. 2021. Multi-modal egocentric activity recognition using multi-kernel learning. *Multimedia Tools and Applications* 80, 11 (2021), 16299–16328.

[4] Martin Azizyan, Ionut Constandache, and Romit Roy Choudhury. 2009. Surroundsense: mobile phone localization via ambience fingerprinting. In *Proceedings of the 15th annual international conference on Mobile computing and networking*. 261–272.

[5] Frederic Charles Bartlett. 1995. *Remembering: A study in experimental and social psychology*. Cambridge university press.

[6] Esma Mansouri Benssassi, Juan-Carlos Gomez, LouAnne E Boyd, Gillian R Hayes, and Juan Ye. 2018. Wearable assistive technologies for autism: opportunities and challenges. *IEEE Pervasive Computing* 17, 2 (2018), 11–21.

[7] Sejal Bhalla, Mayank Goel, and Rushil Khurana. 2021. Imu2doppler: Cross-modal domain adaptation for doppler-based activity recognition using imu data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–20.

[8] Marc Bolanos, Mariella Dimiccoli, and Petia Radeva. 2016. Toward storytelling from visual lifelogging: An overview. *IEEE Transactions on Human-Machine Systems* 47, 1 (2016), 77–90.

[9] Marc Brysbaert and Michaël Stevens. 2018. Power analysis and effect size in mixed effects models: A tutorial. *Journal of cognition* 1, 1 (2018), 9.

[10] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.

[11] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 721–725.

[12] Wenqiang Chen, Jiaxuan Cheng, Leyao Wang, Wei Zhao, and Wojciech Matusik. 2024. Sensor2Text: Enabling Natural Language Interactions for Daily Activity Tracking Using Wearable Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 192 (Nov. 2024), 26 pages. doi:10.1145/3699747

[13] Zhe Chen, Chao Cai, Tianyue Zheng, Jun Luo, Jie Xiong, and Xin Wang. 2021. RF-based human activity recognition using signal adapted convolutional neural network. *IEEE Transactions on Mobile Computing* 22, 1 (2021), 487–499.

[14] Bhawana Chhaglani, Sarmistha Sarna Gomasta, Yuvraj Agarwal, Jeremy Gummeson, and Prashant Shenoy. 2025. FeatureSense: Protecting Speaker Attributes in Always-On Audio Sensing System. *arXiv preprint arXiv:2505.24115* (2025).

[15] Junho Choi, Chang Choi, Hoon Ko, and Pankoo Kim. 2016. Intelligent healthcare service using health lifelog analysis. *Journal of medical systems* 40, 8 (2016), 188.

[16] Caterina Cinel, Cathleen Cortis Mack, and Geoff Ward. 2018. Towards augmented human memory: Retrieval-induced forgetting and retrieval practice in an interactive, end-of-day review. *Journal of Experimental Psychology: General* 147, 5 (2018), 632.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[18] Vladimir Despotovic, Peter Pocta, and Andrej Zgank. 2022. Audio-based Active and Assisted Living: A review of selected applications and future trends. *Computers in Biology and Medicine* 149 (2022), 106027.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[20] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 736–740.

[21] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[22] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. In *International conference on machine learning*. PMLR, 1068–1077.

[23] Daniel A Epstein. 2015. Personal informatics in everyday life. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 429–434.

[24] Irfan A Essa. 2002. Ubiquitous sensing for smart and aware environments. *IEEE personal communications* 7, 5 (2002), 47–49.

[25] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* 9 (2021), 391–409.

[26] Emilio Ferrara. 2024. Large language models for wearable sensor-based human activity recognition, health monitoring, and behavioral modeling: A survey of early trends, datasets, and challenges. *Sensors* 24, 15 (2024), 5045.

[27] Figen Beken Fikri, Kemal Oflazer, and Berrin Yanıkoğlu. 2021. Semantic similarity based evaluation for abstractive news summarization. In *Proceedings of*

the 1st workshop on natural language generation, evaluation, and metrics (GEM 2021). 24–33.

[28] Jessica Forde, Ruochen Zhang, Lintang Sutawika, Alham Aji, Samuel Cahyawijaya, Genta Indra Winata, Minghao Wu, Carsten Eickhoff, Stella Biderman, and Ellie Pavlick. 2024. Re-evaluating evaluation for multilingual summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 19476–19493.

[29] Biying Fu, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. 2020. Sensing technology for human activity recognition: A comprehensive survey. *Ieee Access* 8 (2020), 83791–83820.

[30] Barbara Furletti, Paolo Cintia, Chiara Renso, and Laura Spinsanti. 2013. Inferring human activities from GPS tracks. In *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*. 1–8.

[31] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.

[32] Shayan Gharib, Konstantinos Drossos, Emre Cakir, Dmitriy Serdyuk, and Tuomas Virtanen. 2018. Unsupervised adversarial domain adaptation for acoustic scene classification. *arXiv preprint arXiv:1808.05777* (2018).

[33] Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768* (2024).

[34] Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. 2025. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. *arXiv preprint arXiv:2507.08128* (2025).

[35] Yuan Gong, Yu-An Chung, and James Glass. 2021. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778* (2021).

[36] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. *arXiv preprint arXiv:2305.10790* (2023).

[37] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. 2022. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839* (2022).

[38] Morgan Harvey, Marc Langheinrich, and Geoff Ward. 2016. Remembering through lifelogging: A survey of human memory augmentation. *Pervasive and Mobile Computing* 27 (2016), 14–26.

[39] Steve Hodges, Emma Berry, and Ken Wood. 2011. SenseCam: A wearable camera that stimulates and rehabilitates autobiographical memory. *Memory* 19, 7 (2011), 685–696.

[40] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[41] 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* (2024).

[42] Yasha Iravantchi, Karan Ahuja, Mayank Goel, Chris Harrison, and Alanson Sample. 2021. Privacymic: Utilizing inaudible frequencies for privacy preserving daily activity recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[43] Sijie Ji, Xinzhe Zheng, and Chenshu Wu. 2024. Hargpt: Are llms zero-shot human activity recognizers?. In *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*. IEEE, 38–43.

[44] Hideo Joho, Masaki Matsubara, Norihiko Uda, Chieko Mizoue, and Rahmi Rahmi. 2023. Lifelogging by senior citizens: implications from a light-weight GPS-based study. *F1000Research* 12 (2023), 1461.

[45] Emil Jovanov and Aleksandar Milenkovic. 2011. Body area networks for ubiquitous healthcare applications: opportunities and challenges. *Journal of medical systems* 35, 5 (2011), 1245–1254.

[46] Vaiva Kalnikaite, Abigail Sellen, Steve Whittaker, and David Kirk. 2010. Now let me see where i was: understanding how lifelogs mediate memory. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2045–2054.

[47] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. 2013. A review on video-based human activity recognition. *Computers* 2, 2 (2013), 88–131.

[48] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating Captions for Audios in The Wild. In *NAACL-HLT*.

[49] Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831* (2024).

[50] Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. 2021. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069* (2021).

[51] Maansa Krovvidi, Zhiyi Shi, Sushanta Mohan Rakshit, Anagha Ravi Shankara, Vivek Jain, and Quinn Jacobson. 2025. Activity Recognition using RF and IMU Sensor Data Fusion. In *Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems*. 104–109.

[52] Amel Ksibi, Ala Saleh D Alluhaidan, Amina Salhi, and Sahar A El-Rahman. 2021. Overview of lifelogging: current challenges and advances. *IEEE Access* 9 (2021), 62630–62641.

[53] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications magazine* 48, 9 (2010), 140–150.

[54] Nicholas D Lane, Ye Xu, Hong Lu, Shaohan Hu, Tanzeem Choudhury, Andrew T Campbell, and Feng Zhao. 2011. Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In *Proceedings of the 13th international conference on Ubiquitous computing*. 355–364.

[55] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proceedings of the 31st annual ACM symposium on user interface software and technology*. 213–224.

[56] Eric C Larson, TienJui Lee, Sean Liu, Margaret Rosenfeld, and Shwetak N Patel. 2011. Accurate and privacy preserving cough sensing using a low-cost microphone. In *Proceedings of the 13th international conference on Ubiquitous computing*. 375–384.

[57] Junwoo Lee and Bummo Ahn. 2020. Real-time human action recognition with a low-cost RGB camera and mobile robot platform. *Sensors* 20, 10 (2020), 2886.

[58] Matthew L Lee and Anind K Dey. 2008. Lifelogging memory appliance for people with episodic memory impairment. In *Proceedings of the 10th international conference on Ubiquitous computing*. 44–53.

[59] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[60] Kaylee Yaxuan Li, Yasha Iravantchi, Yichen Zhu, Hyunmin Park, and Alanson P Sample. 2025. HandSAW: Wearable Hand-based Event Recognition via On-Body Surface Acoustic Waves. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 1 (2025), 1–29.

[61] Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804* (2022).

[62] Hong Lu, Wei Pan, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. 2009. Soundsense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*. 165–178.

[63] Lingjuan Lyu, Xuanli He, Yee Wei Law, and Marimuthu Palaniswami. 2017. Privacy-preserving collaborative deep learning with application to human activity recognition. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1219–1228.

[64] Saif Mahmud, Vineet Parikh, Qikang Liang, Ke Li, Ruidong Zhang, Ashwin Ajit, Vipin Gunda, Devansh Agarwal, François Guimbretière, and Cheng Zhang. 2024. ActSonic: recognizing everyday activities from inaudible acoustic wave around the body. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–32.

[65] Mladen Milošević, Michael T Shrove, and Emil Jovanov. 2011. Applications of smartphones for ubiquitous health monitoring and wellbeing management. *Journal of Information Technology and Applications* 1, 1 (2011), 7–15.

[66] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Eunsol Soul Choi, Xuhai Xu, Joanna Kuc, Jeremy F Huckins, Jason Holden, Colin Depp, et al. 2024. Contextual ai journaling: Integrating llm and time series behavioral sensing technology to promote self-reflection and well-being using the mindscape app. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[67] Erik Oudman, Isabelle F Klukas, Tijmen van Teijlingen, and Albert Postma. 2025. First-person and third-person lifelogging improves episodic memory. *Acta Psychologica* 255 (2025), 104929.

[68] Xiaomin Ouyang and Mani Srivastava. 2024. LLMSense: Harnessing LLMs for High-level Reasoning Over Spatiotemporal Sensor Traces. In *2024 IEEE 3rd Workshop on Machine Learning on Edge in Sensor Systems (SenSys-ML)*. IEEE, 9–14.

[69] Huy Phan, Oliver Y Chén, Philipp Koch, Lam Pham, Ian McLoughlin, Alfred Mertins, and Maarten De Vos. 2019. Unifying isolated and overlapping audio event detection with multi-label multi-task convolutional recurrent neural networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 51–55.

[70] Angelica Poli, Annachiara Strazza, Stefania Cecchi, and Susanna Spinsante. 2020. Identification issues associated with the use of wearable accelerometers in lifelogging. In *International Conference on Human-Computer Interaction*. Springer, 338–351.

[71] Stefan Poslad. 2011. *Ubiquitous computing: smart devices, environments and interactions*. John Wiley & Sons.

[72] Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-Yiin Chang, and Tara Sainath. 2019. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing* 13, 2 (2019), 206–219.

[73] Amon Rapp and Federica Cena. 2016. Personal informatics for everyday life: How users without prior self-tracking experience engage with personal data. *International Journal of Human-Computer Studies* 94 (2016), 1–17.

[74] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[75] Ricardo Ribeiro, Alina Trifan, and António JR Neves. 2022. Lifelog Retrieval From Daily Digital Data: Narrative Review. JMIR Mhealth Uhealth 10, 5 (2022), e30517.

[76] Roger C Schank and Robert P Abelson. 2013. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology press.

[77] Abigail J Sellen, Andrew Fogg, Mike Aitken, Steve Hodges, Carsten Rother, and Ken Wood. 2007. Do life-logging technologies support memory for the past? An experimental study using SenseCam. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 81–90.

[78] Mohit Shah, Brian Mears, Chaitali Chakrabarti, and Andreas Spanias. 2012. Lifelogging: Archival and retrieval of continuously recorded audio using wearable devices. In *2012 IEEE International Conference on Emerging Signal Processing Applications*. IEEE, 99–102.

[79] Mostafa Al Masum Shaikh, Md Khademul Islam Molla, and Keikichi Hirose. 2008. Automatic life-logging: A novel approach to sense real-world activities by environmental sound cues and common sense. In *2008 11th International Conference on Computer and Information Technology*. IEEE, 294–299.

[80] Ana Rita Silva, Maria Salomé Pinho, Luís Macedo, Christopher Moulin, Salomé Caldeira, and Horácio Firmino. 2017. It is not only memory: Effects of sensecam on improving well-being in patients with mild alzheimer disease. *International psychogeriatrics* 29, 5 (2017), 741–754.

[81] Sibo Song, Vijay Chandrasekhar, Ngai-Man Cheung, Sanath Narayan, Liyuan Li, and Joo-Hwee Lim. 2014. Activity recognition in egocentric life-logging videos. In *Asian conference on computer vision*. Springer, 445–458.

[82] Ke Sun, Chunyu Xia, Xinyu Zhang, Hao Chen, and Charlie Jianzhong Zhang. 2024. Multimodal daily-life logging in free-living environment using non-visual egocentric sensors on a smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 1 (2024), 1–32.

[83] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289* (2023).

[84] Nirmalya Thakur and Chia Y Han. 2021. A review of assistive technologies for activities of daily living of elderly. *arXiv preprint arXiv:2106.12183* (2021).

[85] Ye Tian, Xiaoyuan Ren, Zihao Wang, Onat Gungor, Xiaofan Yu, and Tajana Rosing. 2025. DailyLLM: Context-Aware Activity Log Generation Using Multi-Modal Sensors and LLMs. *arXiv preprint arXiv:2507.13737* (2025).

[86] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[87] Endel Tulving and Donald M Thomson. 1973. Encoding specificity and retrieval processes in episodic memory. *Psychological review* 80, 5 (1973), 352.

[88] Doménique Van Gennip, Elise Van Den Hoven, and Panos Markopoulos. 2016. The phenomenology of remembered experience: A repertoire for design. In *Proceedings of the European Conference on Cognitive Ergonomics*. 1–8.

[89] Tijmen Van Teijlingen, Erik Oudman, and Albert Postma. 2022. Lifelogging as a rehabilitation tool in patients with amnesia: A narrative literature review on the effect of lifelogging on memory loss. *Neuropsychological Rehabilitation* 32, 10 (2022), 2646–2672.

[90] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern recognition letters* 119 (2019), 3–11.

[91] Wei Wei, Hongning Zhu, Emmanouil Benetos, and Ye Wang. 2020. A-crnn: A domain adaptation model for sound event detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 276–280.

[92] Jacob Westfall, David A Kenny, and Charles M Judd. 2014. Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General* 143, 5 (2014), 2020.

[93] Jason Wu, Chris Harrison, Jeffrey P Bigham, and Gierad Laput. 2020. Automated class discovery and one-shot interactions for acoustic activity recognition. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[94] Huatao Xu, Panrong Tong, Mo Li, and Mani Srivastava. 2024. Autolife: Automatic life journaling with smartphones and llms. *arXiv preprint arXiv:2412.15714* (2024).

[95] Huatao Xu, Pengfei Zhou, Rui Tan, and Mo Li. 2023. Practically adopting human activity recognition. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[96] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 220–233.

[97] Yuxuan Ye, Edwin Simpson, and Raul Santos Rodriguez. 2024. Using similarity to evaluate factual consistency in summaries. *arXiv preprint arXiv:2409.15090* (2024).

[98] Mingfang Zhang, Yifei Huang, Ruicong Liu, and Yoichi Sato. 2024. Masked video and body-worn imu autoencoder for egocentric action recognition. In *European Conference on Computer Vision*. Springer, 312–330.

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

[100] Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A Ali Heydari, Girish Narayanswamy, Maxwell A Xu, Ahmed A Metwally, Shawn Xu, Jake Garrison, Xuhai Xu, et al. 2025. SensorLM: Learning the Language of Wearable Sensors. *arXiv preprint arXiv:2506.09108* (2025).

[101] Haizhong Zheng, Elisa Tsai, Yifu Lu, Jiachen Sun, Brian R Bartoldson, Bhavya Kailkhura, and Atul Prakash. 2024. Elfs: Label-free coreset selection with proxy training dynamics. *arXiv preprint arXiv:2406.04273* (2024).

[102] Ya-Li Zheng, Xiao-Rong Ding, Carmen Chung Yan Poon, Benny Ping Lai Lo, Heye Zhang, Xiao-Lin Zhou, Guang-Zhong Yang, Ni Zhao, and Yuan-Ting Zhang. 2014. Unobtrusive sensing and wearable devices for health informatics. *IEEE transactions on biomedical engineering* 61, 5 (2014), 1538–1554.