# BABBLE: Bridging Structured Labels and Natural Language for Infant-Centric Home Audio Captioning

Anonymous ACL submission

## Abstract

Children's early development is shaped by contingent vocal exchanges with caregivers, yet current audio large language models (LLMs) often fail in infant-centric home recordings because they are trained primarily on adult-directed, lexical speech and coarse web-scale audio. As a result, they struggle with non-lexical vocalizations (e.g., infant babbling and crying) and the fine-grained temporal structure needed to interpret naturalistic caregiver–infant interactions. We introduce **BABBLE**, a compact audio–language modeling framework that bridges structured developmental annotations and natural-language supervision by converting time-stamped labels into captions. BABBLE combines Whisper-derived semantic features with wav2vec 2.0 acoustic representations in a dual-encoder architecture and supports frame-level labeling, event-level prediction, diarization-oriented outputs, and captioning through a unified formulation. Experiments on infant-centric home audio from 63 families (infants aged 3–14 months) with family-disjoint splits show that BABBLE outperforms recent audio LLMs and strong audio-only baselines for speaker and vocalization prediction, improving captioning metrics and reducing diarization error. These results indicate that structured-to-caption supervision is an effective strategy for extending audio–language models to underrepresented, privacy-sensitive, and non-lexical real-world audio domains. Our code are available at `https://anonymous.4open.science/r/BABBLE/`

## 1 Introduction

Children's early development is shaped by everyday interactions with caregivers (Fogel, 1993; National Research Council, 2000; Rosenblum et al., 2019), and contingent vocal exchanges and vocal turn-taking are central modes of communication during the first year of life (Stern et al., 1975; Jaffe et al., 2001; Harder et al., 2015). Unlike adult conversation, infant-caregiver interactions contain infants' non-speech vocalizations (babbling, fussing, crying) alongside caregivers' infant-directed speech and song (motherese, lullabies). These vocalization types differ from adult-directed speech in both acoustics and interaction structure, complicating diarization and audio captioning in naturalistic home recordings of infants and their caregivers. Although infant vocalizations (Zeifman, 2001) and caregivers' infant-directed speech and song (Hilton et al., 2022) exhibit cross-cultural regularities, motivating models that generalize across families and recording conditions, home recordings can also have substantial variability; numbers of family members present in home audio recordings can vary (Sethna et al., 2017), (Oh et al., 2015) as can the presence of other household speakers and background sounds.

These properties challenge common assumptions in contemporary audio understanding models. Developmental datasets are often annotated with highly structured supervision that specifies vocalization types, speaker roles, and fine-grained temporal boundaries. In contrast, modern audio captioning and audio large language models are usually trained using coarse, weakly aligned natural-language (Goel et al., 2025; Ghosh et al., 2025; Kong et al., 2024; Xu et al., 2025; Chu et al., 2024; Tang et al., 2023). This mismatch between the supervision that is available in infant research and the objectives used to train captioning models makes it difficult to transfer models trained on adult-centric or web-scale audio to infant home environments.

Given this supervision mismatch, it is natural to ask whether recent audio language models can be applied to infant home recordings. While these models perform well on general-purpose benchmarks, they are largely shaped by adult-directed, lexical speech and relatively coarse acoustic events found in web-scale data and natural-language supervision. Infant-centric home audio instead con-

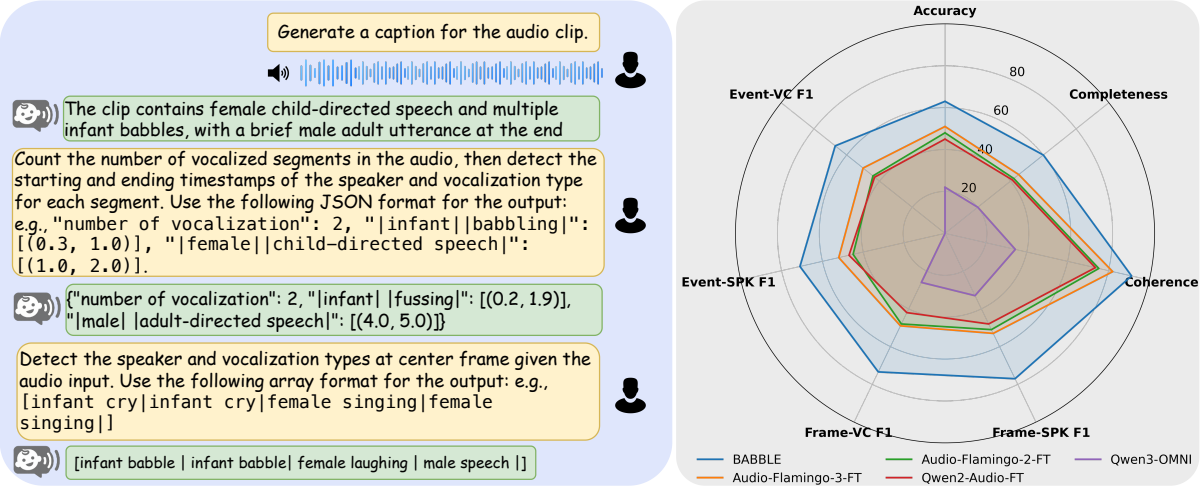



Figure 1: Overview of BABBLE, with three different outputs: a) frame-based speaker and vocalization prediction with high reliability b) event-based speaker diarization c) audio captioning for natural understanding. It outperforms off-the-shelf finetuned audio-LLM, demonstrating its effectiveness in infant-centric real-world environments.

tains frequent non-lexical vocalizations, multiple speakers, short-duration events, and fine-grained temporal structure. Thus, off-the-shelf audio LLMs often degrade sharply under non-lexical events.

We address the supervision mismatch in infant home audio captioning while avoiding a key practical bottleneck, reliance on word-level transcripts. Transcription is often infeasible or undesirable in this setting because infant vocalizations are largely non-lexical, which makes automatic speech recognition unreliable, and because transcripts can increase privacy and data-sharing constraints for sensitive in-home recordings. We introduce BABBLE, a compact 2.6B-parameter audio language model (audio-LM) trained using natural-language captions generated from structured developmental annotations by a large language model. Concretely, we convert fine-grained labels that capture speaker roles, vocalization types, and temporal boundaries into captions, producing caption supervision that preserves domain-specific structure while aligning with standard language-model training objectives.

BABBLE uses a dual-encoder audio front end because infant home audio mixes weak lexical content with rich non-lexical cues such as cries, babbles, and prosody, and models must capture both semantic content and fine-grained acoustics. The model combines Whisper features, which are effective for speech and higher-level semantics, with wav2vec 2.0 features, which better preserve low-level acoustic and temporal detail. In addition to caption generation, BABBLE supports 0.1-second frame-level prediction with 2 to 10-second context, event-level aggregation over short windows, speaker diarization, and speaker count estimation. This design allows a single model to produce high-level descriptions while retaining fine-grained structure, and we evaluate how these auxiliary outputs affect robustness in multi-speaker infant–caregiver interactions.

**Contributions.** This work makes three contributions. First, we introduce a structured-to-caption supervision strategy that converts high-resolution annotations of infant-centric home audio into natural-language summaries, enabling caption-style learning without word-level transcription. Second, we propose BABBLE, a compact audio language model that jointly supports frame-level labeling, event-level prediction, and diarization-oriented outputs, enabling modeling of short events and overlapping speakers in home recordings. Third, using infant-centric audio from 63 families with family-disjoint splits, we show that BABBLE consistently outperforms recent open audio language models and strong audio-only baselines on captioning metrics, fine-grained speaker and vocalization labeling, and diarization error rate, supported by both automatic rubric scoring and a complementary human evaluation.

## 2 Related Works

**Child-Centered Family Audio Analysis.** Limitations of the LENA system's coarse speaker labels and moderate accuracy in home environments (Xu et al., 2014; Cristia et al., 2021) have motivated supervised learning approaches for child-centered audio analysis, including parent–infant diarization (Xie et al., 2019; Cristia et al., 2018), infant cry

Table 1: Distribution of annotated durations (minutes) across partitions and task categories.

| Partition | #Families | CHN | | | | FAN | | | | MAN | | CXN | OVL | SEC | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | BAB | CRY | FUS | LAU | CDS | ADS | LAU | SNG | CDS | ADS | | | FAN | MAN |
| Train | 48 | 129.7 | 49.2 | 102.9 | 3 | 121.5 | 195.7 | 8.7 | 43.2 | 89 | 83.2 | 114.2 | 82 | 8.4 | 8.8 |
| Dev | 5 | 9.1 | 0.6 | 8.9 | 0.1 | 7.1 | 5.6 | 0.2 | 2.5 | 0.6 | 3.9 | 1.5 | 3.3 | 1.3 | 0.04 |
| Test | 10 | 24.2 | 2.5 | 13.2 | 0.8 | 20.4 | 12.3 | 0.9 | 2.6 | 1.9 | 3.4 | 5.6 | 2.9 | 1.5 | 0 |

detection (Yao et al., 2022), and vocalization classification (Li et al., 2021). Recently, speech foundation models such as wav2vec 2.0 (Baevski et al., 2020) and Whisper (Radford et al., 2022) have improved representation learning under limited supervision; in particular, child-centered pretraining (e.g., wav2vec2-LL4300) and self-supervised models have shown clear gains over adult-centric pretraining for family audio analysis and speaker diarization (Li et al., 2023a; Charlot et al., 2025; Fan et al., 2025; Xu et al., 2024).

**Audio Language Models and Audio Captioning.** Recent audio language models combine pretrained audio encoders with generative language models to support a range of speech and audio understanding tasks (Gong et al., 2023b; Chu et al., 2023; Ji et al., 2024; Huang et al., 2024; Elizalde et al., 2023). While prior work has studied their capabilities and limitations (Sakshi et al., 2025; yu Huang et al., 2024), most evaluations focus on utterance-level classification or coarse audio-text alignment, leaving fine-grained temporal, paralinguistic, and vocalization-level understanding relatively underexplored (Wang et al., 2023b, 2024). In parallel, audio captioning research has progressed from dataset-driven approaches such as AudioCaps and Clotho (Kim et al., 2019; Drossos et al., 2020) to LLM-based systems including SALMONN, Qwen-Audio, SpeechGPT, and Audio-Flamingo variants (Tang et al., 2023; Chu et al., 2023; Zhang et al., 2023a; Ghosh et al., 2025; Goel et al., 2025). Trained largely on web-scale audio–text data dominated by adult speech and environmental sounds, these models achieve strong performance on standard benchmarks but typically rely on coarse, weakly aligned supervision.

**Positioning of Our Work.** Prior work on child-centered family audio has largely focused on diarization, cry detection, and vocalization classification, whereas audio captioning and audio LLMs are typically trained on web-scale audio–text pairs dominated by adult speech and environmental sounds. In infant home recordings, supervision more often takes the form of time-stamped developmental annotations (speaker roles, vocalization types, event boundaries) rather than transcripts or natural captions. We connect these lines by training BABBLE with caption supervision generated from structured developmental annotations using a large language model, enabling caption-style learning without transcription. Unlike web-scale keyword-to-caption augmentation from coarse tags, our captions are generated from fine-grained labels and preserve speaker-role and temporal structure for infant-centric, multi-speaker audio.

## 3 Data

### 3.1 Data Collection

We collect naturalistic, child-centered audio recordings from family homes to study infant vocal interactions under everyday conditions. The cohort includes recordings from **63 families**, with infants aged 3–14 months at the time of recording. Recordings capture typical household activity, including infant vocalizations and caregiver speech, as well as speech from other household members and background sounds that vary across homes. Audio is recorded using infant-centric wearable audio recording platforms LittleBeats (Islam et al., 2024) worn or placed in the home according to the study protocol detailed in Appendix B.1. To evaluate cross-household generalization, we organize the dataset at the family level and define splits that are disjoint by family (details in Section 3.2). Section 8 describes ethical considerations, IRB, consent, and data privacy.

### 3.2 Annotation and Segmentation

To annotate LB home recordings, we segment each day-long recording into 2-minute clips and prioritize clips likely to contain infant and mother vocalizations. Two trained annotators label time-bounded vocalizations in Praat (Boersma, 2006) across multiple tiers (infant, female adult, male adult, sibling). From each 2-minute clip, we extract overlapping windows (2–30s; stride 1s) for training; for evaluation, we use non-overlapping windows to avoid correlated test samples. For frame-based evaluation, we score predictions at

0.1s resolution across the full segment. Additionally, to evaluate robustness under controlled multi-speaker conditions, we construct a synthetic benchmark by inserting secondary-speaker speech into real 2-minute clips. The details of synthetic data generation are described in Appendix B.6.

Infant vocalizations are labeled as BAB (babbling), FUS (fussing), and CRY (crying), and caregiver vocalizations as ADS (adult-directed speech), CDS (child-directed speech), SNG (singing), and LAU (laughter). We annotate sibling speech as CXN and use FAN/MAN tiers when multiple female/male adults are present. Primary speakers are CHN/FAN/MAN/CXN, and non-caregiver adults are labeled as secondary (SEC-FAN/SEC-MAN).

We use family-disjoint train/dev/test splits of 48/5/10 families (Table 1). Annotated durations are computed per speaker-role tier from time-stamped labels; because tiers can overlap, these totals do not equal the unique recording duration. Chunk counts per task are reported in Appendix B.3.

**Annotation Quality.** Each annotated clip was independently reviewed by a second annotator who verified label assignments and temporal boundaries, with particular attention to short events and overlapping vocalizations. Agreement was computed at 0.1s resolution over speaker and vocalization labels, and clips were accepted if $\kappa \geq 0.7$ or diagonal mismatch $\leq$ 3s (Appendix B.2).

### 3.3 Caption Construction

In addition to structured, time-stamped annotations, we curate natural-language captions to support caption-based training and evaluation without requiring word-level transcripts. We generate captions for all extracted windows using GPT-5.1-mini as a structured caption generator conditioned on symbolic event annotations, yielding 378k training captions and 30k evaluation captions.

**Inputs and Format.** Each captioned instance corresponds to *2 to 30-second* acoustic window. For every window, we construct a JSON-style annotation that specifies the number of active vocal sources and their semantic categories (e.g., infant babbling, child-directed speech, background speech), together with their temporal extents.

Each tuple specifies the speaker role, event type, and time span within the window (Figure 5). The captioning goal is to map this symbolic record to a concise description of the window's vocal content without explicitly mentioning timestamps.

**LLM-based Conversion and Constraints.** We prompt GPT-5.1-mini to generate a single-sentence caption with (i) lowercase formatting, (ii) dataset-style phrasing, and (iii) faithful coverage of the annotated vocalizations. To encourage semantic abstraction and improve generalization across temporal variations, we omit temporal intervals so captions describe *what* occurs rather than *when*; for windows with no *annotated* vocal activity, the model emits a canonical "no vocalizations" caption. These captions provide weak supervision for downstream audio–language training; Appendix B.4 details the prompt and caption-length statistics.

**Released Artifacts.** We do not release raw audio or annotations due to in-home privacy constraints. We will release model checkpoints, training/inference code, and the caption-generation templates and validation scripts. Example JSON–caption pairs are shown in Figure 5 (Appendix B.4).

## 4 BABBLE

BABBLE is an audio language model for infant-centric home recordings. Given raw audio and a text prompt, it extracts audio features using two complementary encoders (Whisper and wav2vec 2.0), aligns each encoder stream to the LM embedding space using lightweight CNN projection modules, and concatenates the aligned audio tokens with tokenized text inputs. A language decoder (TinyLlama-1.1B-chat) then generates the final output. Figure 2 overviews the architecture.

### 4.1 Audio Encoder

We fuse wav2vec 2.0 and Whisper-large-v2 to capture both fine-grained acoustics (short non-lexical events and precise timing) and higher-level semantic/prosodic cues for robust audio captioning and auxiliary prediction.

**Wav2vec 2.0 Encoder.** As the acoustic branch, we use wav2vec2-LL4300 (Li et al., 2023a), pretrained on 4,300h of unlabeled child-centered audio and fine-tuned end-to-end on labeled data (Appendix E). For each frame, we extract all-layer hidden states ($D$=768), compute a trainable layer-weighted average, and pass the sequence to the alignment module, supporting short non-lexical events and fine-grained frame-level prediction.

**Whisper Encoder.** We use the Whisper-large-v2 encoder (Radford et al., 2022) as the semantic branch. We adapt it to infant-centric audio in an audio-only setup (Appendix E) and use final-layer encoder representations. We find it most stable for downstream generation and captioning.

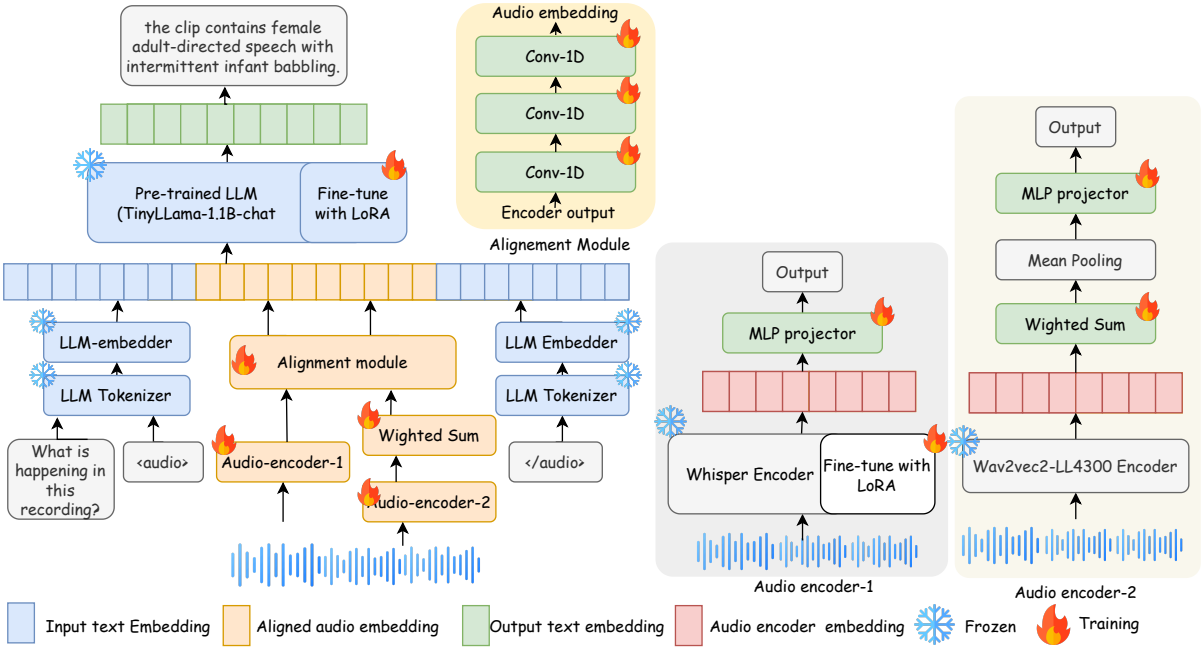



Figure 2: Overview of the BABBLE architecture. Raw audio is processed by two complementary audio encoders (Whisper and wav2vec 2.0), whose outputs are independently aligned to the language model embedding space using lightweight convolutional alignment modules. The aligned audio embeddings are concatenated with tokenized text inputs and passed to a pre-trained language model (TinyLlama-1.1B-chat), which is adapted using LoRA.

This branch provides semantically rich features that complement wav2vec 2.0's fine-grained acoustics.

### 4.2 Alignment Module

Let $\mathbf{H}^{(w)} \in \mathbb{R}^{B\times T_w\times D_w}$ and $\mathbf{H}^{(s)} \in \mathbb{R}^{B\times T_s\times D_s}$ denote wav2vec 2.0 and Whisper encoder outputs, respectively. Because they differ in temporal resolution and feature dimension, we align each stream independently into a shared space compatible with the language decoder. The alignment module is a three-layer 1-D convolutional network (CNN) operating along the temporal dimension. Given an encoder output $\mathbf{H} \in \mathbb{R}^{B\times T\times D}$, it produces aligned tokens $\hat{\mathbf{Z}} \in \mathbb{R}^{B\times T'\times D_{\text{LLM}}}$, where $D_{\text{LLM}}$ is TinyLlama's embedding dimension. All layers use kernel size 2 to capture short-range temporal dependencies while preserving fine temporal structure. Unless noted otherwise we use stride 1 with padding to preserve resolution ($T' = T$). The alignment is defined as $\hat{\mathbf{Z}} = \text{Conv}_3(\sigma(\text{Conv}_2(\sigma(\text{Conv}_1(\mathbf{H})))))$. Here, $\text{Conv}_1$, $\text{Conv}_2$, and $\text{Conv}_3$ are 1D convolutional layers and $\sigma$ is the ReLU activation.

This convolutional alignment (i) projects heterogeneous encoder representations into the language-model embedding space and (ii) injects local temporal context for robustness to short events and minor timing shifts. By aggregating neighboring frames, it is more stable than per-frame linear projection, and it is far more parameter-efficient than attention-based aligners (e.g., Q-Former) without adding cross-modal attention, which we found unnecessary for our low-dimensional outputs. Overall, the three-layer CNN offers a strong capacity–stability trade-off, consistently beating linear projection while remaining lighter than Q-Former (Ablation 5.5).

### 4.3 Feature Fusion

After alignment, we fuse the two audio streams to form a single audio token sequence. Let $\hat{\mathbf{Z}}^{(w)} \in \mathbb{R}^{B\times T'_w\times D_{\text{LLM}}}$ and $\hat{\mathbf{Z}}^{(s)} \in \mathbb{R}^{B\times T'_s\times D_{\text{LLM}}}$ be the aligned wav2vec 2.0 and Whisper outputs. We concatenate them along the temporal dimension:

$$\hat{\mathbf{Z}}_a = \left[\hat{\mathbf{Z}}^{(w)}; \hat{\mathbf{Z}}^{(s)}\right]_T \in \mathbb{R}^{B\times(T'_w+T'_s)\times D_{\text{LLM}}} \quad (1)$$

This avoids explicit resampling between encoder time grids while still letting the decoder attend to both streams. The resulting sequence $\hat{\mathbf{Z}}_a$ is the final audio representation for all tasks.

### 4.4 Language Decoder and Training Objective

We use TinyLlama-1.1B-chat (Zhang et al., 2024) as the language decoder. It takes as input the concatenation of fused audio tokens $\hat{\mathbf{Z}}_a$ and tokenized text (instruction/query), and generates outputs autoregressively. We train with a standard causal language modeling objective, minimizing the negative log-likelihood over generated tokens, where each token is predicted conditioned on the full input sequence and previously generated token.

$$\mathcal{L}_{\text{gen}} = -\sum_{i=0}^{N-1} \log P\left(z_t^i \mid \mathbf{Z}_t^{<i}, \mathbf{Z}_s\right), \quad (2)$$

### 4.5 Training Strategy

We adopt a multi-stage training strategy to stabilize multimodal optimization. We use Adam optimizer and apply LoRA to selected TinyLlama modules (rank 16 and 32), with learning rates of $2 \times 10^{-4}$ for the alignment modules, $1 \times 10^{-5}$ for encoder fine-tuning, and $2 \times 10^{-4}$ for TinyLlama LoRA parameters, plus a 3% linear warmup.

Training proceeds in three stages: (1) train only the alignment modules while freezing the audio encoders and TinyLlama; (2) fine-tune wav2vec 2.0 and Whisper while continuing to train the alignment modules at a reduced learning rate, jointly refining acoustic representations and alignment for infant-centric audio; and (3) freeze the encoders and alignment modules and train only TinyLlama LoRA parameters to adapt generation while preserving the learned audio representations.

### 4.6 Tasks and Output Formats

We train BABBLE to support three tasks to produce structural fine-grained predictions and caption.

**Frame-Based Prediction.** Given 2–30s of audio, BABBLE predicts labels at 0.1s resolution for primary speaker (SPK), vocalization type (VC), and secondary speaker (SEC). For SPK, BABBLE predicts silence, 4 primary speakers (Table 1), and overlap. For VC, it predicts individual vocalization types for four primary speakers and reports the average (AVG VC). For SEC, the model predicts SIL, SEC_FAN, SEC_MAN, and SEC_OVL.

**Event-Based Prediction.** For 2–30s windows, BABBLE outputs event-level SPK/VC/SEC using the same inventories and predicts the number of active speakers (speaker count).

**Audio Captioning.** For a 2–30s segment, it generates a natural-language summary of salient speaker roles and vocalization patterns (Section 3.5).

## 5 Result and Discussion

### 5.1 Evaluation Protocol

**Metrics.** Frame- and event-based tasks are evaluated with unweighted F1 for SPK/VC/SEC, along with diarization error rate (DER) and mean absolute error (MAE) for speaker count. For audio captioning, since standard captioning metrics are poorly aligned with infant-centric audio, we use GPT-5.1 as an automatic judge and report *Completeness*, *Coherence*, and *Accuracy*. The judge receives the model output and ground-truth structured events (not raw audio) and scores each criterion under a fixed rubric (Appendix F.7). Because BABBLE is generative, we apply deterministic normalization/validation before scoring (label normalization, timestamp correction/clipping for event outputs, and inventory and temporal-order checks) for all metrics. Invalid/discarded outputs receive no credit. Decoding details and retention rates are in Appendix B.5. We also conduct a *human caption evaluation* with three independent raters.

**Baselines.** We compare BABBLE to two baseline families: *audio–language models* and *audio-only models*. For audio–language models, we evaluate recent open-source systems (Audio-Flamingo-2 (Ghosh et al., 2025), Audio-Flamingo-3 (Goel et al., 2025), Qwen3-Omni (Xu et al., 2025), and Qwen2-Audio (Chu et al., 2024)) in both *zero-shot* form and, when feasible, after *fine-tuning* on our data. To isolate BABBLE's contributions beyond using an LM decoder, we also evaluate strong audio-only backbones and adaptation strategies, including Whisper-AT (Gong et al., 2023a), wav2vec2-LL4300 (Li et al., 2023a), BS-SSAMBA (Fan et al., 2025), Whisper with LoRA fine-tuning, and a Whisper–wav2vec 2.0 feature-combination baseline. All audio-only baselines use the same downstream evaluation protocol as BABBLE for frame- and event-based tasks.

### 5.2 Comparison with Audio-LLMs

Table 2 compares BABBLE with open-source audio–language baselines under the unified evaluation protocol in Section 5.1, using identical inputs, prompts, and post-processing for all methods.

**Audio Captioning.** BABBLE substantially outperforms both zero-shot and fine-tuned audio-LLMs by up to **+45.8** and **+12.2**, with the largest gains in *accuracy* and *completeness*. This indicates better coverage of the role- and event-structure present in infant-centric windows rather than generic, speech-dominant descriptions.

**Frame-based Prediction.** BABBLE achieves higher SPK (up to 30.3%) and VC (32.42%) performance than fine-tuned audio-LLMs, with higher agreement (Cohen's $\kappa$ 41.32-48.8%), suggesting improved recovery of fine-grained, role-specific temporal structure at 0.1s resolution.

**Event-based Prediction and Diarization.** BABBLE improves event-level SPK/VC by 26.3/25.2 on average, reduces DER by *28.1%* relative to fine-tuned audio-LLMs, and yields lower speaker-count MAE of *0.8*, indicating more accurate event extents in time in addition to correct labels.

Table 2: Comparison of BABBLE with open-source audio–language models on audio captioning, frame-based classification, and event-based prediction. * indicates models fine-tuned on our infant-centric data; others are zero-shot. BABBLE achieves the best performance across all tasks.

| | Audio Captioning | | | | Frame-based | | | | Event-based | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | Comp. | Coher. | Avg. | Avg. SPK | | Avg. VC | | Avg. SPK | | Avg. VC | | Diarization |
| | | | | | F1-score | Kappa | F1-score | Kappa | F1-score | Kappa | F1-score | Kappa | DER (%) |
| Audio-Flamingo-2 | 17.4 | 13.1 | 26.4 | 19.0 | 24.4 | 9.3 | 18.9 | -4.3 | - | - | - | - | - |
| Audio-Flamingo-3 | 21.7 | 19.4 | 34.7 | 25.3 | 31.4 | 12.8 | 23.3 | 5.1 | - | - | - | - | - |
| Qwen3-OMNI | 22.7 | 20.5 | 34.4 | 25.9 | 33.5 | 19.5 | 26.3 | 9.2 | - | - | - | - | - |
| Qwen2-Audio | 14.3 | 13.9 | 27.1 | 18.4 | 17.2 | 3.2 | 17.5 | 0 | - | - | - | - | - |
| Audio-Flamingo-2* | 48.8 | 43.1 | 75.3 | 55.7 | 51.6 | 42.2 | 48.6 | 33.5 | 45.9 | 37.6 | 44.1 | 34.3 | 57.4 |
| Audio-Flamingo-3* | 51.4 | 45.1 | 82.1 | 59.5 | 53.7 | 45.3 | 49.6 | 35.2 | 52.7 | 45.1 | 50.2 | 39.6 | 51.7 |
| Qwen2-Audio* | 45.6 | 41.3 | 73.6 | 53.2 | 48.1 | 34.3 | 42.4 | 24.0 | 47.6 | 33.1 | 43.1 | 31.7 | 58.1 |
| **BABBLE** | **63.3** | **60.5** | **91.3** | **71.7** | **77.1** | **77.2** | **73.4** | **67.8** | **71.5** | **71.8** | **67.5** | **60.4** | **28.1** |

Table 3: Comparison of BABBLE with audio-only encoder baselines for frame- and event-based prediction. Results are reported per primary speaker (CHN/FAN/MAN/CXN) and for SEC speakers. BABBLE outperforms all audio-only baselines across tiers, improving F1 by 7.8–19.25% on average and reducing DER to 28.1%.

| | Farme-based | | | | | Event-based | | | | | Diarization | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CHN | FAN | MAN | CXN | SEC | CHN | FAN | MAN | CXN | SEC | Spk. (MAE) | DER (%) |
| Whisper-AT | 58.4 | 69.6 | 72.7 | 77.2 | 48.2 | 51.4 | 52.1 | 48.3 | 70.4 | 46.5 | 1.9 | 38.9 |
| Wav2Vec2-LL4300 | 60.2 | 64.2 | 60 | 77.3 | 34.7 | 53.3 | 48.5 | 46.2 | 70.1 | 31.3 | 2.1 | 41.1 |
| BS SSAMBA | 58.1 | 65.1 | 61.6 | 72.4 | - | - | - | - | - | - | - | - |
| Whisper w/ LoRA | 57.3 | 66.1 | 61.4 | 74.1 | 45.1 | 51.2 | 54.2 | 54.3 | 68.1 | 45.2 | 2.3 | 45.1 |
| Whisper-wav2vec2 | 60.1 | 68.3 | 65.4 | 77.4 | 52.5 | 56.4 | 59.1 | 56.5 | 71.6 | 48.3 | 1.7 | 34.5 |
| **BABBLE-post** | **67.1** | **72.9** | **72.7** | **80.5** | **61.2** | 63.1 | 65.6 | 61.1 | 74.3 | 55.2 | 1.1 | 29.7 |
| **BABBLE** | **67.1** | **72.9** | **72.7** | **80.5** | **61.2** | **65.3** | **66.3** | **62.4** | **75.8** | **57.2** | **0.8** | **28.1** |

[1] BABBLE-post uses the same post-processing method as audio-only models for the Event-based result calculation.

Across tasks, improvements align with three design choices: (i) *infant-centric caption supervision* derived from structured role/event annotations, capturing non-lexical and overlapping vocalizations underrepresented in web-scale audio–text; (ii) a *merged encoder front end* combining fine timing cues with richer semantic features; and (iii) *CNN alignment* that adds local temporal context while mapping both streams into the decoder space. With identical prompts, constrained schema, and post-processing across methods, differences reflect modeling rather than scoring.

### 5.3 Comparison with audio-only models

Table 3 compares BABBLE with strong audio-only baselines under the same frame- and event-based protocol (Section 4.6). Audio-only encoders are fine-tuned for *frame-level* prediction; for *event-level*, diarization, and speaker count, we convert frame predictions to events using the same deterministic post-processing for all audio-only methods (Appendix F). We report BABBLE end-to-end and **BABBLE-post**, which applies the same post-processing as the audio-only baselines.

BABBLE achieves the best overall performance across speaker roles and vocalization tiers: it improves SEC (52.5 to 61.2) over the strongest Whisper–wav2vec2 baseline, leads on event-based average F1 (67.5), and yields better diarization and counting (DER 28.1%, MAE 0.8). BABBLE-post is slightly lower, as expected since the matched pipeline discards some of BABBLE's structured event output, but enables a controlled comparison. Even with this slightly lower performance BABBLE outforms all of the audio-only baselines. We attribute gains to infant-centric supervision and joint audio-LM, with a dual-encoder (wav2vec2 and Whisper) that preserves short non-lexical cues and role structure in infant-centric home audio.

### 5.4 Human Evaluation

We performed a human evaluation on a random subset of 500 captions generated by BABBLE. Three independent evaluators rated each caption for **accuracy** on a **10-point Likert scale** by comparing each caption against (i) the **structured event annotations** for the window and (ii) the **reference caption** generated from those annotations using GPT-5.1-mini (Section 3.3). Evaluators were **blind to model identity** and did not see GPT-5.1 scores. Average scores were **7.1/10**, **6.8/10**, and **6.4/10**. These human ratings are consistent with the trends observed under our GPT-5 judging protocol, providing complementary evidence that the automatic evaluation captures human-perceived faithfulness on infant-centric captions.

Table 4: Audio-captioning ablation across audio encoders: the merged wav2vec 2.0+Whisper front end consistently improves caption *accuracy*, *completeness*, and *coherence* over using either encoder alone.

| Audio-Encoder | Acc. | Comp. | Coher. | Avg. |
|---|---|---|---|---|
| Wav2Vec 2.0 | 55.7 | 51.9 | 83.1 | 63.6 |
| Whisper w/ LoRA | 58.7 | 53.2 | 87.3 | 66.4 |
| Wav2Vec 2.0 + Whisper | 63.3 | 60.5 | 91.3 | 71.7 |

Table 5: Audio-captioning performance decreases as input length increases: short segments achieve higher scores than longer windows.

| Length | Acc. | Comp. | Coher. | Avg. |
|---|---|---|---|---|
| 2s | 65.4 | 63.3 | 93.6 | 74.1 |
| 5s | 65.7 | 63.9 | 93.3 | 74.3 |
| 10s | 61.2 | 58.5 | 89.4 | 69.7 |
| 30s | 58.2 | 55.9 | 86.1 | 66.7 |

Table 6: Audio captioning performance under different numbers of training stages shows that multi-stage training improves accuracy, completeness, and overall caption quality compared to single-stage training.

| # Training-stage | Acc. | Comp. | Coher. | Avg. |
|---|---|---|---|---|
| 1 | 59.4 | 55.7 | 89.1 | 68.1 |
| 3 | 63.3 | 60.5 | 91.3 | 71.7 |

### 5.5 Ablation Study

We conduct controlled ablations on **audio captioning**, which best reflects BABBLE's ability to summarize fine-grained acoustics into high-level descriptions; frame- and event-based ablations are deferred to Appendix C. All variants use the same data, caption supervision, decoding/post-processing, and the same GPT-5.1 judging protocol on the same split. We ablate (*i*) the audio encoder, (*ii*) input length, (*iii*) training strategy, and (*iv*) the alignment module.

**Effect of Audio Encoder** Table 4 compares audio encoders for captioning. Wav2vec2.0 alone captures fine-grained acoustics but provides weaker semantic cues, yielding moderate scores. Whisper with LoRA improves all metrics, consistent with its semantically richer representations. The merged wav2vec2.0+Whisper front end performs best overall, with the largest gains in completeness and coherence, indicating complementary contributions from fine temporal detail (wav2vec 2.0) and higher-level semantic cues (Whisper).

**Effect of Audio Length** Table 5 shows caption quality peaks for short inputs (2–5s) and drops for longer windows (10–30s). We therefore report results across 2–30s, since the required context is often unknown a priori. Performance degrades as longer segments contain more events, overlaps, and speaker turns to compress into one sentence, increasing omissions, role confusion, and reduced coherence. This highlights the difficulty of long-context captioning in dense infant home audio and motivates auxiliary structure predictions (see Appendix C for additional ablation).

**Effect of Training Strategy** Table 6 demonstrates that the three-stage training strategy consistently improves captioning performance over single-stage training. Progressively optimizing alignment, then refining acoustic representations, and finally adapting the language model stabilizes learning and improves generation faithfulness and coverage, yielding higher-quality captions across all metrics.

Table 7: Alignment ablation (linear, Q-Former, CNN) for audio captioning, showing the benefit of lightweight temporal modeling with CNN alignment.

| Alignment Module | Acc. | Comp. | Coher. | Avg. | #Param |
|---|---|---|---|---|---|
| Linear | 58.8 | 54.1 | 87.0 | 66.6 | 11.1M |
| Q-Former | 61.3 | 58.6 | 89.4 | 69.8 | 60M |
| CNN | 63.3 | 60.5 | 91.3 | 71.7 | 29.4M |

**Effect of alignment module.** Table 7 ablates alignment while keeping the encoders and language model fixed. A three-layer MLP performs worst, suggesting limited ability to reconcile heterogeneous encoder features. Q-Former (Li et al., 2023b) improves scores with attention-based alignment but adds parameters and complexity (configuration in Appendix D.2). Our CNN alignment performs best across metrics, indicating that lightweight temporal aggregation is an effective and parameter-efficient way to handle short events and minor timing shifts in infant home audio.

## 6 Conclusion

We presented BABBLE, a compact audio–language model for infant-centric home recordings that must handle non-lexical vocalizations, multiple speakers, and fine-grained temporal structure. BABBLE uses structured-to-caption supervision, converting high-resolution developmental role and event annotations into natural-language captions, which enables caption-style learning without word-level transcription and mitigates in-home privacy constraints. The proposed dual-encoder front end (wav2vec 2.0 for fine-grained acoustics and Whisper for semantic cues) with CNN-based alignment yields consistent gains over open-source audio–LLMs and strong audio-only baselines on captioning, frame-based and event-based prediction, and diarization. Our results suggest that aligning structured developmental supervision with natural-language generation is a practical path for adapting audio foundation models to underrepresented, real-world, and sensitive environments where raw data sharing and transcription are infeasible.

## 7 Limitations

Although BABBLE demonstrates exceptional performance on the target speaker labeling task in real-world naturalistic home recordings, the model's accuracy in audio captioning degrades as the input segment length increases, particularly beyond 10 seconds. This suggests that the current architecture struggles to aggregate and summarize high-density acoustic information in longer windows, where multiple overlapping speakers and varied vocalization types are more frequent.

Furthermore, while our approach avoids the need for privacy-sensitive word-level transcripts, it still requires high-resolution, time-stamped symbolic annotations for training. The labor-intensive nature of this manual coding remains a bottleneck for scaling the system to larger, uncurated datasets. We only evaluated a single lightweight LLM model; exploring LLM models with larger parameters may enhance their reasoning capabilities and yield improved performance.

Finally, the model is susceptible to hallucinations and formatting errors common in small-scale language models when model temperature is set close to 1, necessitating external post-processing and time-enforcement strategies to ensure valid and temporally consistent outputs.

**Future Work.** To address the performance degradation observed in longer recordings, we plan to explore a new architecture that can effectively summarize day-long home audio without losing fine-grained temporal detail. Additionally, we aim to integrate BABBLE with generalized, web-scale audio foundation models to enhance its acoustic diversity and zero-shot capabilities. By combining our specialized infant-centric dual-encoder with broader audio-text representations, the model could better distinguish between relevant caregiver speech and complex, overlapping household background sounds.

Furthermore, we intend to incorporate synchronized ECG and IMU data to integrate physiological and motion signals, which will allow the model to generate more holistic summaries of infant well-being, such as correlating crying events with heart rate variability or physical activity levels. Finally, we will scale the framework to larger backbones, such as 7B or 13B parameter LLMs, to evaluate if increased model capacity can further reduce hallucinations and improve the coherence of complex multi-speaker interaction summaries.

## 8 Ethical Consideration

This study collects naturalistic in-home audio from infants and family members under IRB approval at the host institution with caregiver-informed consent and IRB-approved compensation. All data files are marked with identification numbers only and stored on secure password-protected institutional servers. Data are only accessible by trained study personnel, and data are not shared outside the research team. Parents are also able to turn off the recording device at any time. Parents are also informed that the majority of their audio recordings are processed automatically without human intervention and that human coders only review small samples of the audio. Because in-home audio may contain sensitive or identifying speech, we do not produce or store word-level transcripts; instead, we use structured developmental annotations (speaker roles, vocalization types, event boundaries) and derive caption supervision from these annotations.

## 9 Risk Statement

Open-sourcing our model allows the community to evaluate its performance on their own recordings within the intended use case. However, misuse of BABBLE could lead to unintended or untested outputs, though it poses minimal risk for malicious activities. Additionally, the interpretability of our model remains limited, making it difficult to identify failure cases. We recommend careful deployment of such systems with human oversight, ongoing auditing of training data sources, and fu- ture work on explainability and robust alignment to reduce these risks.

## References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Takanori Ashihara, Takafumi Moriya, Shota Horiguchi, Junyi Peng, Tsubasa Ochiai, Marc Delcroix, Kohei Matsuura, and Hiroshi Sato. 2024. Investigation of speaker representation for target-speaker speech processing. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 423–430. IEEE.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th International*

*Conference on Neural Information Processing Systems (NeurIPS)*, pages 12449–12460.

Subrata Biswas, Mohammad Nur Hossain Khan, and Bashima Islam. 2025a. Owl: Geometry-aware spatial reasoning for audio large language models. *arXiv preprint arXiv:2509.26140*.

Subrata Biswas, Mohammad Nur Hossain Khan, and Bashima Islam. 2025b. Raven: Query-guided representation alignment for question answering over audio, video, embedded sensors, and natural language. *arXiv preprint arXiv:2505.17114*.

Paul Boersma. 2006. Praat: Doing phonetics by computer. http://www.praat.org/. Accessed: 2026-01-04.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE.

Théo Charlot, Tarek Kunze, Maxime Poli, Alejandrina Cristia, Emmanuel Dupoux, and Marvin Lavechin. 2025. Babyhubert: Multilingual self-supervised learning for segmenting speakers in child-centered long-form recordings. *arXiv preprint arXiv:2509.15001*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Alejandrina Cristia, Shobhana Ganesh, Marisa Casillas, and Sriram Ganapathy. 2018. Talker diarization in the wild: The case of child-centered daylong audio-recordings. In *Proceedings of Interspeech*, pages 2583–2587.

Alejandrina Cristia, Marvin Lavechin, Camila Scaff, Melanie Soderstrom, Caroline Rowland, Okko Räsänen, John Bunce, and Elika Bergelson. 2021. A thorough evaluation of the language environment analysis (lena) system. *Behavior Research Methods*, 53:467–486.

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)*, pages 736–740. IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Xulin Fan, Jialu Li, Mark Hasegawa-Johnson, and Nancy L McElwain. 2025. Band-split self-supervised mamba for infant-centered audio analysis. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2795–2799.

Alan Fogel. 1993. *Developing through Relationships: Origins of Communication, Self, and Culture*. University of Chicago Press.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and 1 others. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*.

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023a. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. *arXiv preprint arXiv:2307.03183*.

Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023b. Joint audio and speech understanding. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng

Gao, and Xiangyu Yue. 2024. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595.

Susanne Harder, Theis Lange, Gitte F. Hansen, Mette Væver, and Simo Køppe. 2015. A longitudinal study of coordination in mother–infant vocal interaction from age 4 to 10 months. *Developmental Psychology*, 51(12):1778–1790.

C. B. Hilton, C. J. Moser, M. Bertolo, H. Lee-Rubin, D. Amir, C. M. Bainbridge, J. Simson, D. Knox, L. Glowacki, E. Alenuma, A. Galbarczyk, G. Jasienska, C. T. Ross, M. Beth Neff, A. Martin, L. K. Cirelli, S. E. Trehub, J. Song, M. Kim, and S. A. Mehr. 2022. Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, and 1 others. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23802–23804.

Mingyue Huo, Abhinav Jain, Cong Phuoc Huynh, Fanjie Kong, Pichao Wang, Zhu Liu, and Vimal Bhat. 2025. Beyond speaker identity: Text-guided target speech extraction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Bashima Islam, Nancy L. McElwain, Jialu Li, Maria I. Davila, Yannan Hu, Kexin Hu, Jordan M. Bodway, Ashutosh Dhekne, Romit Roy Choudhury, and Mark Hasegawa-Johnson. 2024. Preliminary technical validation of littlebeats: A multimodal sensing platform to capture cardiac physiology, motion, and vocalizations. *Sensors*, 24(3):901.

Joseph Jaffe, Beatrice Beebe, Stanley Feldstein, Cynthia L. Crown, and Michael D. Jasnow. 2001. Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the Society for Research in Child Development*, 66(2):1–132.

Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, and 1 others. 2024. Wavchat: A survey of spoken dialogue models. *Preprint*, arXiv:2411.13577.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Yidi Jiang, Zhengyang Chen, Ruijie Tao, Liqun Deng, Yanmin Qian, and Haizhou Li. 2024b. Prompt-driven target speech diarization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11086–11090. IEEE.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. 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)*, pages 119–132.

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*.

Marvin Lavechin, Ruben Bousbib, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia. 2020. An open-source voice type classifier for child-centered daylong recordings. In *Proceedings of Interspeech*.

Jialu Li, Mark Hasegawa-Johnson, and Nancy L. McElwain. 2021. Analysis of acoustic and voice quality features for the classification of infant and mother vocalizations. *Speech Communication*, 133:41–61.

Jialu Li, Mark Hasegawa-Johnson, and Nancy L. McElwain. 2023a. Towards robust family-infant audio analysis based on unsupervised pretraining of wav2vec 2.0 on large-scale unlabeled family audio. In *Proceedings of Interspeech*, pages 1035–1039.

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

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023c. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

National Research Council. 2000. *From Neurons to Neighborhoods: The Science of Early Childhood Development*. National Academies Press, Washington, DC.

Wonjung Oh, Brenda L. Volling, and Richard Gonzalez. 2015. Trajectories of children's social interactions with their infant sibling in the first year: A multidimensional approach. *Journal of Family Psychology*, 29(1):119–129.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Katherine L. Rosenblum, Carolyn J. Dayton, and Maria Muzik. 2019. Infant social and emotional development: Emerging competence in a relational context. In Charles H. Zeanah, editor, *Handbook of Infant Mental Health*, pages 95–119. The Guilford Press.

S. Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *International Conference on Learning Representations (ICLR)*.

Vaheshta Sethna, Eleanor Perry, Jo Domoney, Jonathan Iles, Lamprini Psychogiou, Nicola E. L. Rowbotham, Alan Stein, Lynne Murray, and Paul G. Ramchandani. 2017. Father–child interactions at 3 months and 24 months: Contributions to children's cognitive development at 24 months. *Infant Mental Health Journal*, 38(3):378–390.

Daniel N. Stern, Joseph Jaffe, Beatrice Beebe, and Samuel L. Bennett. 1975. Vocalizing in unison and in alternation: Two modes of communication within the mother–infant dyad. *Annals of the New York Academy of Sciences*, 263:89–100.

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*.

Yunlong Tang, Daiki Shimada, Jing Bi, Mingqian Feng, Hang Hua, and Chenliang Xu. 2024. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. *arXiv preprint arXiv:2403.16276*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hualei Wang, Jianguo Mao, Zhifang Guo, Jiarui Wan, Hong Liu, and Xiangdong Wang. 2024. Leveraging language model capabilities for sound event detection. In *Proceedings of Interspeech*, pages 4803–4807.

Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. 2023a. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv preprint arXiv:2304.14407*.

Yingzhi Wang, Mirco Ravanelli, and Alva Yacoubi. 2023b. Speech emotion diarization: Which emotion appears when? In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Jiamin Xie, Leibny Paola Garcia-Perera, Daniel Povey, and Sanjeev Khudanpur. 2019. Multi-plda diarization on children's speech. In *Proceedings of Interspeech*, pages 376–380.

Anfeng Xu, Kevin Huang, Tiantian Feng, Lue Shen, Helen Tager-Flusberg, and Shrikanth Narayanan. 2024. Exploring speech foundation models for speaker diarization in child–adult dyadic interactions. In *Proceedings of Interspeech*, pages 5193–5197.

D. Xu, Jeffrey A. Richards, and J. Gilkerson. 2014. Automated analysis of child phonetic production using naturalistic recordings. *Journal of Speech, Language, and Hearing Research*, 57(5):1638–1650.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen3-omni technical report. *arXiv preprint arXiv:2503.20215*.

Xuewen Yao, Megan Micheletti, Mckensey Johnson, Edison Thomaz, and Kaya de Barbaro. 2022. Infant crying detection in real-world environments. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE.

Chien yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, and 1 others. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE.

Debra M. Zeifman. 2001. An ethological analysis of human infant crying: Answering tinbergen's four questions. *Developmental Psychobiology*, 39.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *Preprint*, arXiv:2401.02385.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Katerina Zmolikova, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Černocký, and Dong Yu. 2023. Neural target speech extraction: An overview. *IEEE Signal Processing Magazine*, 40(3):8–29.

# A More Related Works

This section includes additional models, datasets, and encoder variants relevant to our work that were not cited in the related work of the main paper due to space constraints. We list them here for completeness and to acknowledge recent progress in MLLMs and audio-LLMs.

**Language Models.** Mixtral (Jiang et al., 2024a), LLaMA-2(Touvron et al., 2023), LLaMA-3(Dubey et al., 2024), Phi (Abdin et al., 2024), Qwen2 (Team et al., 2024), PaLM (Chowdhery et al., 2023), OPT (Zhang et al., 2022).

**Audio/Visual MLLMs.** Video-LLava (Lin et al., 2023), VideoChat (Li et al., 2023c), VideoChat-GPT (Maaz et al., 2023), ChatVideo (Wang et al., 2023a), AViCuna (Tang et al., 2024), Video-LLaMA2 (Cheng et al., 2024), BEATs (**?**), OneLLM (Han et al., 2024), RAVEN (Biswas et al., 2025b), Video-LLaMA (Zhang et al., 2023b), OWL (Biswas et al., 2025a).

**Target and role differentiation in multi-speaker mixtures.** A related challenge to our work is identifying a target speaker or target role in multi-speaker recordings. Traditional approaches often rely on extra cues such as enrolled speaker profiles, visual features such as lip movement, or spatial direction to guide separation or identification (Ashihara et al., 2024; Zmolikova et al., 2023). More recently, text-guided target speech extraction has been explored as a flexible way to specify the desired speaker using text input (Jiang et al., 2024b; Huo et al., 2025). However, these methods are rarely studied in naturalistic home environments with frequent overlap and diverse background sounds. In the family-audio setting, (Lavechin et al., 2020) differentiated primary speakers from unknown sources by adding an additional class without relying on explicit speaker cues.

# B Dataset Curation and Annotation

## B.1 Data Collection Protocol and Setup

The LittleBeats (LB) system simultaneously records three data streams—electrocardiogram (ECG), motion, and audio—while participants engage in their normal daily activities. To capture in-home data, participating families were instructed to complete day-long recording sessions (approximately 8 hours per day). All adults present in the home during recording periods (e.g., parents, grandparents, or caregivers) provided informed consent via a secure, web-based form administered by the research team. In instances where non-consenting adults were present, parents were instructed to disable the device for the duration of their presence. Families were compensated with a USD $100 electronic gift card for their participation. During data collection, the LB device was placed in a pocket of a wearable garment worn by the infant. Figure 3 illustrates the LB device and the in-home data collection setup. Details of the data collection setup are described in (Islam et al., 2024).

Table 8: Cohen's kappa scores for human annotators (test dataset only)

| | SPK | SEC | CHN | FAN | MAN | CXN |
|---|---|---|---|---|---|---|
| Version 1 | 0.88 | 0.86 | 0.62 | 0.76 | 0.70 | 0.79 |
| Version 2 | 0.88 | 1.00 | 0.85 | 0.92 | 0.86 | 0.83 |

## B.2 Inter-Coder Reliability Assessment

Inter-annotator reliability was calculated across two versions: in **version 1**, two annotators worked independently on labeling speakers and vocalization types. For files that reached a Cohen's kappa score $\geq 0.7$ or an off-diagonal sum of misclassifications $\leq 3$ seconds), annotator 1's files were designated as the final version. In **version 2**, for files that did not meet version 1 agreement criteria, both annotators independently revised their annotations after reviewing each other's version 1 files, resulting in version 2. A second reliability assessment was then conducted. If version 2 met the above agreement criteria, annotator 1's file was designated as the final version. Otherwise, the remaining disagreements were discussed, and a consensus file was created and designated the final version. Table 8 shows the inter-coder reliability for annotations of these two versions on test files.

## B.3 Number of Data for Each Task

We segment every two-minute annotated audio segment into 2 to 30-second chunks for each frame-based, event-based, and captioning task. We create overlapping segments with a stride of 1s for training and development, and non-overlapping segments for event-based test data. However, for frame-based data, we generate overlapping chunks with a stride of 0.1s to predict every 0.1s resolution. Table 9 shows the number of audio segments for each task and for each duration.


Figure 3: (A) LittleBeats device case; (B) LittleBeats supplies, including electrocardiogram leads, electrodes, charger, and shirt; and (C) an infant wearing LittleBeats at home.

Table 9: Data distribution for frame-based, event-based, and captioning tasks.

| Audio length | Frame-based | | Event-based | | Captioning | |
|---|---|---|---|---|---|---|
| | Train | Test | Train | Test | Train | Test |
| 2s | 343910 | 90440 | 103173 | 13680 | 103173 | 13680 |
| 5s | 335240 | 88160 | 100572 | 5472 | 100572 | 5472 |
| 10s | 320790 | 84360 | 96237 | 2736 | 96237 | 2736 |
| 30s | 262990 | 69160 | 78897 | 912 | 78897 | 912 |

### B.4 GPT Prompt for Caption Generation

We use GPT-5.1-mini to generate captions from structured event-based labels. Additionally, we also created a variety of questions for each caption to introduce variability in the questions asked to BABBLE-LM. The prompt to generate captions and questions for 5s audio is shown in Figure 4. The captions are generated in a batched manner to reduce API costs. Examples of event-based annotations to generate captions are shown in Figure 5.

### B.5 Output post-processing and validity checks.

Due to hallucinations and spelling errors, LLM output is not always valid. Thus, we use post-processing to decode the correct output for frame-based and event-based results. Before scoring, we apply deterministic normalization and validation: (i) *Label normalization:* map common spelling variants to the closest valid inventory entry and correct speaker-vocalization pairs jointly when needed (e.g., "inant brying" $\rightarrow$ "infant crying"). (ii) *Timestamp constraints (event-based):* clip timestamps to $[0, T]$ and swap endpoints if start time exceeds end time. (iii) *Inventory and ordering checks:* discard outputs that are not in the defined inventory or violate temporal ordering after correction. Discarded/invalid outputs are counted as incorrect for all metrics (i.e., they do not receive credit). However, we find that the retention rate for frame-based output is 99.7% and for event-based prediction is 99.1%. This shows that our model learns all output formats with high accuracy.

### B.6 Synthetic Data Construction

To evaluate robustness under controlled multi-speaker conditions, we construct a synthetic benchmark by inserting secondary-speaker speech into real 2-minute clips. For each recording $R$ from a family $F$ in the training, development, or test split, we randomly select a secondary speaker $T_{\text{sec}}$ (FAN or MAN) from a different family $F' \neq F$, using speech segments (ADS or CDS) drawn from a different recording $R' \in F'$. Synthetic clips are generated within each split using only source material from families assigned to that split, preserving a leave-one-family-out scheme and preventing leakage of speaker characteristics across training and evaluation sets. For each base recording, we repeat this process three times using three independently sampled secondary speakers ($|T_{\text{sec}}| = 3$), resulting in a synthetic dataset that is three times larger than the original set of 2-minute clips. Secondary-speaker segments are inserted at random timestamps, with a maximum total inserted duration of 10 seconds per 2-minute clip (about 20% of the clip duration). The algorithm of synthetic audio generation is described in Algorithm 1

## C Additional Ablation Study

In this section, we provide the performance of BABBLE-LM on frame-based, event-based, and diarization tasks for different design choices.

### C.1 Audio-encoder Effect

Table 10 compares BABBLE-LM with three audio encoder configurations across frame-based classification, event-based detection, and speaker diarization. The combined Wav2Vec2+Whisper encoder consistently outperforms single-model baselines

```
"""
 Build a single prompt for a batch of lines. batch_items: list of (line_index, annotation_line)
"""
header = """
You are preparing training data for an audio caption + question dataset.

You will receive multiple annotation lines. each line describes a 5-second audio clip
with fields like:
{  "number of vocalization": 2,   "|infant| |babbling|": [4.6,5.0],   "|female| |adult-directed speech|":
[0.0,5.0], }

meaning:
- "number of vocalization" = number of distinct vocal sources / speakers in the 0–5 second clip.
- entries like "|speaker| |event|" specify what kind of vocalization occurs (infant babbling, crying,
  adult-directed speech, child-directed speech, singing, etc.) and at which times.
- speakers may include: infant, female, male, child, irrelevant female, irrelevant male.
- "irrelevant" means background speech not directed to the target child.

For EACH annotation line, you must:
1. Create a short audio caption in dataset style:
   - lowercase
   - single sentence
   - natural language description of what happens in the clip (what voices / vocalizations are present).
   - Do NOT mention timestamps, but integrate them in a natural language description.

2. Create a short generic question about the clip:
   - also lowercase
   - generic, not referring to specific timestamps.
   - ask about what is heard in the clip (e.g., "what happens in this audio clip?",
     "what sounds are present in this five-second clip?", etc.), and vary the wording across items.

Return your answer as a **JSON array only**, no extra text, no markdown.
each element of the array MUST have this structure:

{
  "line_index": <integer>,   // the line_index I gave you
  "caption": "the clip contains ...",
  "question": "what ... ?"
}
```

Figure 4: Caption generation prompt using GPT-5.1-mini from structured event-based annotation.

| Event-based Structured Annotation | GPT-5.1-mini Generated Caption |
|---|---|
| { "number of vocalization": 4, "\|infant\| \|babbling\|": [1.2,1.7], "\|child\| \|speech\|": [1.7,3.0], "\|child\| \|speech\|": [4.2,4.8], "\|irrelevant female\| \|speech\|": [3.4,5.0], } | the clip contains infant babbling and a child is talking, with additional background female speech. |
| { "number of vocalization": 3, "\|female\| \|child-directed speech\|": [0.0,0.9], "\|female\| \|adult-directed speech\|": [1.7,3.1], "\|female\| \|adult-directed speech\|": [3.9,4.8], } | a woman is talking to a child, followed by two instances of female adult-directed speech |
| { "number of vocalization": 3, "\|infant\| \|fussing\|": [0.0,0.5], "\|infant\| \|babbling\|": [1.3,4.7], "\|irrelevant female\| \|speech\|": [0.0,4.3], } | A infant is fussing and babbling while female is talking in the background. |
| { "number of vocalization": 3, "\|infant\| \|babbling\|": [0.5,0.8], "\|female\| \|child-directed speech\|": [2.6,3.1], "\|female\| \|laughter\|": [0.8,1.3], } | the clip contains infant babbling with child-directed female speech followed by a female laughter. |

Figure 5: Example of event-based annotations to generate caption.

across nearly all sound event categories in both frame-based and event-based evaluations, indicating complementary representations captured by the two models. Notably, this fusion yields substantial gains in challenging categories such as SEC and MAN, while also achieving the lowest speaker count error and a markedly reduced DER. These results suggest that integrating self-supervised acoustic representations with large-scale speech models improves robustness across both fine-grained temporal labeling and speaker-level reasoning tasks.

### C.2 Effect of Audio Length

Table 11 examines the effect of input audio length on system performance. For frame-based tasks, performance generally improves with longer segments, peaking around 10s–30s, likely due to increased contextual information. In contrast, event-based performance degrades for longer segments, particularly beyond 10s, suggesting that shorter windows better preserve event localization accuracy. Diarization performance shows a clear trade-off, with shorter segments achieving lower DER and more accurate speaker counts, while longer segments introduce increased speaker confusion. These trends highlight the importance of task-specific temporal granularity when selecting audio segment lengths.

### C.3 Effect of Alignment Module

Table 12 evaluates the impact of different alignment modules for bridging acoustic representations with the language model. Linear alignment yields the weakest performance across all tasks, indicating limited capacity to model temporal and semantic mismatches between modalities. QFormer improves performance substantially by enabling learned cross-modal interactions, but remains inferior to the CNN-based alignment. The CNN alignment module consistently delivers the highest scores across both frame-based and event-based categories, while also achieving the lowest speaker count error and DER. These results highlight the importance of temporally aware convolutional alignment for effective multimodal fusion in continuous audio understanding tasks.

### C.4 Effect of Training Strategy

Table 13 studies the effect of training strategy by comparing single-stage and three-stage optimization schemes. Introducing multi-stage training leads to consistent gains across all sound event categories in both frame-based and event-based evaluations. Notably, three-stage training yields marked improvements in challenging vocalization types such as SEC and MAN, while also reducing speaker count error and DER. These results suggest that progressively refining the model through staged training improves representation alignment and task generalization across diverse audio understanding objectives.

## D Baselines

### D.1 Audio-LLM baselines

**Audio-Flamingo-2.** Audio-Flamingo-2 extends the Flamingo multimodal framework to audio-language modeling by coupling a pretrained audio encoder with a frozen large language model through cross-attention layers. The model relies on lightweight multimodal adapters for training, enabling efficient adaptation to audio understanding tasks such as classification, captioning, and reasoning while preserving the linguistic capabilities of the underlying language model.

**Audio-Flamingo-3.** Audio-Flamingo-3 builds upon Audio-Flamingo-2 by incorporating improved audio representations and more effective cross-modal alignment mechanisms. The model demonstrates stronger performance on temporally structured audio tasks, benefiting from enhanced fusion between acoustic features and language representations, and is designed to better scale to longer audio contexts and more complex audio-language reasoning scenarios.

**Qwen2-Audio.** Qwen2-Audio is an audio-language model that integrates a high-capacity audio encoder with the Qwen family of large language models. Trained on large-scale paired audio-text data, it supports a wide range of audio understanding tasks, including sound event recognition, speech understanding, and audio captioning, with an emphasis on robust generalization across diverse acoustic conditions.

**Qwen3-Omni.** Qwen3-Omni is a unified multimodal foundation model designed to process audio, vision, and text within a single architecture. By jointly modeling multiple modalities, Qwen3-Omni enables cross-modal reasoning and flexible input–output configurations, allowing audio signals to be interpreted either independently or in conjunction with other modalities for complex multimodal understanding tasks.

Table 10: Performance comparison of BABBLE-LM with different audio encoders on frame-based, event-based, and diarization tasks. Combining wav2vec2 and Whisper yields the best performance across all tasks and vocalization tiers.

| Audio Encoder | Farme-based | | | | | Event-based | | | | | Diarization | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CHN | FAN | MAN | CXN | SEC | CHN | FAN | MAN | CXN | SEC | Speaker | DER (%) |
| Wav2Vec2 | 60.1 | 66.3 | 62.6 | 74.5 | 34.5 | 54.1 | 51.1 | 52.7 | 72.1 | 37.5 | 2.1 | 37.2 |
| Whisper w/ LoRA | 58.1 | 68.3 | 62.5 | 76.5 | 39.4 | 52.5 | 56.1 | 56.3 | 71.3 | 41.4 | 2.3 | 38.1 |
| Wav2Vec2+Whisper | 67.1 | 72.9 | 72.7 | 80.5 | 61.2 | 65.3 | 66.3 | 62.4 | 75.8 | 57.2 | 0.8 | 28.1 |

Table 11: Impact of audio segment length on frame-based, event-based, and diarization performance. Results (F1-score) are reported for segment durations ranging from 2 s to 30 s across sound event categories and diarization metrics.

| Audio Length | Farme-based | | | | | Event-based | | | | | Diarization | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CHN | FAN | MAN | CXN | SEC | CHN | FAN | MAN | CXN | SEC | Speaker | DER (%) |
| 2s | 64.2 | 67.3 | 70.4 | 74.5 | 55.6 | 66.5 | 68.4 | 66.5 | 77.3 | 62.1 | 0.3 | 23.1 |
| 5s | 65.2 | 71.2 | 70.3 | 77.4 | 58.5 | 65.1 | 68.7 | 65.4 | 77.1 | 61.9 | 0.7 | 25.9 |
| 10s | 68.1 | 73.4 | 73.1 | 80.9 | 62.3 | 62.4 | 65.4 | 60.7 | 74.3 | 54.2 | 1 | 30.1 |
| 30s | 67.3 | 72.7 | 71.9 | 79.7 | 60.7 | 61.1 | 65.2 | 60.6 | 73.1 | 55.3 | 1.2 | 33.1 |

### D.2 Audio-only Models

**Whisper-AT.** Whisper-AT (Gong et al., 2023a) is an audio model derived from Whisper that is adapted for audio tagging and event-level audio understanding. By leveraging Whisper's pretrained speech representations and fine-tuning them for non-speech audio events, Whisper-AT provides a strong baseline for sound event recognition and audio classification tasks, particularly in scenarios involving speech-dominant acoustic environments.

**Band-Split SSAMBA** Fan et al. propose Band-Split SSAMBA, a self-supervised audio representation learning framework designed for infant-centered home audio analysis. The model extends Self-Supervised Audio Mamba by decomposing spectrogram inputs into frequency subbands using band-specific projections, while employing a shared Mamba state-space encoder to model temporal dependencies. This design improves data efficiency and scalability for long, naturalistic recordings with limited labeled data, and serves as a strong baseline for infant vocalization classification and speaker diarization in home environments.

**Wav2Vec2-LL4300** Li et al. propose an unsupervised pretraining framework for family-infant audio analysis based on wav2vec 2.0, where large-scale unlabeled day-long home recordings are used to learn domain-specific acoustic representations. The pretrained model is fine-tuned for downstream tasks, including parent-infant speaker diarization and vocalization classification, demonstrating that in-domain self-supervised pretraining on home audio substantially reduces domain mismatch compared to models pretrained on general-purpose speech corpora. Their work provides a strong task-specific baseline for family audio understanding in naturalistic home environments.

**Q-Former Configuration.** We employs Q-Former as described in BLIP-2 (Li et al., 2023b) which bridges frozen modality encoders and the frozen large language model. The Q-Former consists of a stack of Transformer layers initialized from a pretrained BERT model. It comprises 12 Transformer layers with 12 self-attention heads per layer and a hidden dimension of 768, along with a feed-forward network of dimension 2048. The model operates on a fixed set of 32 learnable query tokens that interact through self-attention and attend to frozen encoder representations via cross-attention blocks in each layer. The resulting query embeddings are linearly projected into the language model embedding space and provided as prefix tokens, enabling effective cross-modal alignment while keeping both the audio encoder and the language model frozen during training.

Table 12: Performance comparison of BABBLE-LM with different alignment modules on frame-based, event-based, and diarization tasks. The CNN-based alignment module consistently achieves the best performance across all vocalization categories and diarization metrics.

| Alignment Module | Farme-based | | | | | Event-based | | | | | Diarization | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CHN | FAN | MAN | CXN | SEC | CHN | FAN | MAN | CXN | SEC | Speaker | DER (%) |
| Linear | 56.3 | 64.1 | 63.4 | 70.1 | 33.9 | 52.1 | 48.5 | 49.2 | 68.6 | 35.1 | 2.6 | 43.1 |
| QFormer | 63.2 | 69.2 | 65.1 | 75.6 | 52.4 | 55.1 | 61.2 | 62.1 | 73.6 | 51.2 | 1.3 | 33.5 |
| CNN | 67.1 | 72.9 | 72.7 | 80.5 | 61.2 | 65.3 | 66.3 | 62.4 | 75.8 | 57.2 | 0.8 | 28.1 |

Table 13: Effect of the number of training stages in BABBLE-LM implies that multi-stage training significantly improves performance across all tasks and vocalization tiers.

| No. of Training Stages | Farme-based | | | | | Event-based | | | | | Diarization | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CHN | FAN | MAN | CXN | SEC | CHN | FAN | MAN | CXN | SEC | Speaker | DER (%) |
| I | 62.4 | 68.1 | 67.5 | 77.5 | 56.4 | 60.1 | 61.3 | 58.7 | 74.5 | 54.1 | 1.2 | 33.1 |
| III | 67.1 | 72.9 | 72.7 | 80.5 | 61.2 | 65.3 | 66.3 | 62.4 | 75.8 | 57.2 | 0.8 | 28.1 |

# E Open-sourced Model Fine-Tuning Details

## E.1 Audio-LLM Fine-tuning

We fine-tune open-sourced audio-LLM (Audio-Flamingo-2, Audio-Flamingo-3, Qwen-audio-2) using supervised instruction tuning on our paired audio–text data, where each training example consists of a raw audio clip, a natural-language question, and a target caption or response. Audio inputs are processed using the model's native audio encoder and integrated into the language model through Audio-Flamingo's cross-modal fusion mechanism. Training samples are formatted using the official chat template, which interleaves text prompts with audio placeholders, allowing the model to jointly attend to acoustic and linguistic context. To ensure memory efficiency and stable optimization, we adopt parameter-efficient fine-tuning via Low-Rank Adaptation (LoRA), inserting trainable rank-decomposition matrices into the transformer's projection layers while keeping the base model weights frozen. The model is optimized using cross-entropy loss over the assistant tokens only, with gradients accumulated across multiple steps to support large effective batch sizes. Fine-tuning is performed in distributed data-parallel (DDP) mode across multiple GPUs using mixed-precision training (bfloat16), and checkpoints are periodically saved for evaluation and recovery. We use HuggingFace () trainer and model from HuggingFace for fine-tuning purposes. However, we do not fine-tune qwen-omni due to its large size (30B) and resource constraints.

## E.2 Audio-only Models Fine-tuning

**Whisper-AT** We use a similar architecture and hyperparameters as described by the author in the paper. We use the same dataset and the same test-train split as described in BABBLE-LM for frame-based results. Event-based results and time duration for each vocalization are generated using post-processing of the frame-based results.

**Band-split SSAMBA** We do not fine-tune SSAMBA on our dataset; rather, we use the reported result from the paper for comparison. Therefore, we do not have the event-based result for SSAMBA.

**Whisper w/ LoRA** As we train the whisper-encoder with the alignment module and LLM, we fine-tune a whisper-large-v2 on our dataset with LoRA to reduce training cost and avoid multiple training. We keep the base whisper-encoder frozen and use LoRA to the query and value projections of the Whisper encoder, using rank $r = 4$ and scaling factor $\alpha = 8$. We add a feed-forward network to project the output to the appropriate classification task. The model is trained for 20 epochs with the Adam optimizer at a learning rate of 0.001.

**Wav2Vec2-LL4300** We fine-tune wav2vec2-LL4300 with our dataset for frame-based tasks. Raw LB audio is encoded via wav2vec2-LL4300. The encoded audio, represented by 12 transformer layers with a hidden size of 768, is fed into a downstream network. This network consists of

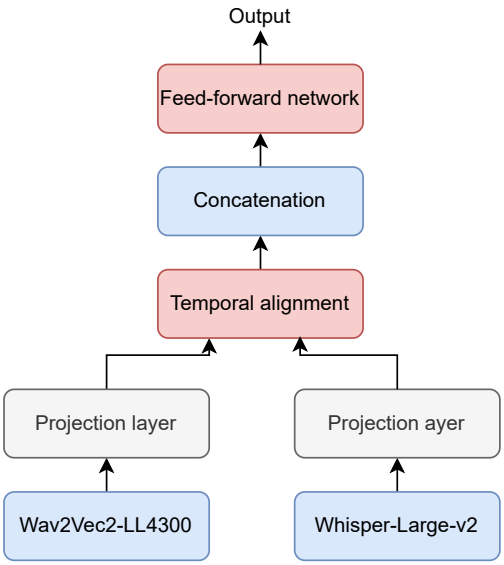



Figure 6: The network architecture of merged Whisper and wav2vec2.

a weighted sum, mean pooling, and a feed-forward network to produce a fixed-dimensional feature of size 384 for classification. We process the audio segments to predict the centered 0.1-second frame. Similar to the other audio model, we post-process the frame-based output to generate the event-based result.

**Merged Wav2vec2 and Whisper** After fine-tuning the wav2vec2 and whisper with our dataset, we combine representations from wav2vec 2.0 and Whisper for audio-only training. To enable fusion, the output of each encoder is first passed through a lightweight feed-forward projection network that maps the original feature dimensions into a shared embedding space. The projected representations are then temporally aligned by resampling one sequence to match the other, ensuring correspondence across time steps. Once aligned, the two streams are concatenated along the feature dimension and optionally passed through an additional feed-forward layer to produce a unified audio representation. This projection-and-concatenation strategy allows the model to jointly leverage complementary low-level acoustic cues and higher-level semantic information. The model is trained for 20 epochs with the Adam optimizer at a learning rate of 0.001. Figure 6 shows the architecture of the merged encoder for audio fine-tuning.

Table 14: Hyperparameters used for training BABBLE-LM, including optimization settings, LoRA configuration, and stage-wise epoch schedule.

| Description | Value |
|---|---|
| Encoder LR | 0.00001 |
| Projector LR | 0.0002 |
| LLM Backbone LR | 0.0002 |
| Epochs | Stage I: 3<br>Stage II: 3<br>Stage III: 3 |
| Warmup ratio | 3 % |
| Embedding size | 2048 |
| LLM Backbone | LLaMA TinyChat-1.1B |
| LoRA Rank | 16 |
| LoRA Alpha | 32 |
| GPU | 4 x H200 (140GB) |
| Global Batch Size | 256 |
| Optimizer | AdamW |

# F Training and Evaluation Details

## F.1 Hyperparameter for Training

BABBLE has 2.6B parameters, including all the encoders, alignment module, and LLM backbone. All hyperparameters used are shown in Table 14.

## F.2 Hardware Requirement

We train our model using four NVIDIA H200 GPUs (140 GB each) with a total CPU memory of 256GB. Training runs for 48 hours with a local batch size of 64 and a global batch size of 256. Evaluation is performed on one NVIDIA H200 GPUs (140 GB) with a batch size of 128.

## F.3 Optimization

Training uses AdamW with a learning rate of $210^{-4}$, weight decay 0.01, and gradient clipping 1.0.

## F.4 Event-based inference from frame-level predictions

To obtain event-based predictions with explicit time durations from the audio-only models, we post-process the outputs of the frame-based classifiers. The frame-based models generate predictions at a fixed temporal resolution of 0.1 s. Consecutive frame-level predictions are merged into a single event as long as adjacent frames correspond to the same speaker and vocalization category. The temporal extent of

```
{"role": "system",
 "content": "You are an intelligent assistant designed to evaluate the quality of generated audio captions.
Your task is to compare a predicted audio caption with the corresponding ground-truth description and assess
how well the prediction matches the reference."
}

{ "role": "user",
  "content": "Please evaluate the following audio caption pair:\n\n
Ground-Truth Caption: {ground_truth}\n
Predicted Caption: {prediction}\n\n
Evaluate the predicted caption based on the following criteria:\n
1. Accuracy: How factually correct the predicted caption is with respect to the ground-truth events (e.g.,
speakers, vocalization types, interactions).\n
2. Completeness: How well the predicted caption covers all salient events present in the ground-truth caption.\n
3. Coherence: How fluent, consistent, and well-structured the predicted caption is as a natural-language
description.\n\n
Provide your evaluation as a Python dictionary string with the following keys:\n
- 'accuracy': a float value between 0 and 10\n
- 'completeness': a float value between 0 and 10\n
- 'coherence': a float value between 0 and 10\n\n
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION.
Only output the Python dictionary string.\n\n
For example:\n {'accuracy': 7.8, 'completeness': 7.2, 'coherence': 8.1}"}
```

Figure 7: System and user prompt used to evaluate the generated caption quality and correctness.

each event is defined by the start time of the first frame and the end time of the last frame in the merged sequence. To suppress spurious short detections, if a merged segment has a duration shorter than 0.3 s, it is merged with the preceding event rather than treated as an independent event. For example, given frame-level predictions $\{\text{FAN}, \text{FAN}, \text{FAN}, \text{FAN}, \text{CHN}, \text{CHN}, \text{FAN}, \text{FAN}\}$ at 0.1 s resolution, the post-processing yields a single event-level output $|\text{female}|$ spanning $[0.0, 0.8]$ s, since the brief infant segment does not meet the minimum duration threshold.

### F.5 Human Evaluator Details

We recruited three annotators through internal advertisements at the host institution. All three annotators are male and aged between 25–35 years old and had a basic understanding of large language models. Participation was voluntary, and no financial incentive was provided.

### F.6 Evaluation Metric

Frame- and event-based tasks are evaluated using unweighted F1 for SPK/VC/SEC classification. For VC, we compute vocalization classification separately for each of the four primary speakers (CHN, FAN, MAN, CXN) and we also report the average classification accuracy (AVG VC). For SEC, we detect secondary-speaker presence with four classes: SIL, SEC_FAN, SEC_MAN, and SEC_OVL. For event-based windows, speaker count is evaluated using mean absolute error (MAE). We additionally report diarization error rate (DER) to quantify diarization quality with a 0.25 s collar and without skipping overlap, computed using pyannote (Bredin et al., 2020).

### F.7 GPT-5.1 Evaluation Prompt

For **audio captioning**, we evaluate captions using **GPT-5.1** as an automatic judge and report three criteria: **Completeness** (coverage of salient events), **Coherence** (fluency and consistency), and **Accuracy** (faithfulness to the ground-truth annotations). Figure 7 depicts the evaluation prompt.

## G Effect of Temperature.

To generate output from BABBLE, we use a temperature setting of 0.2 with $top_p = 0.95$ for more deterministic output. These settings also produce a lower discard rate for event-based and frame-based output. On the other hand, if we increase the tem-

perature to 0.7, the retention rate drops to 97% for frame-based and 92% for event-based output due to not following the output format. However, these settings do not impact the quality of caption generation.

## H Qualitative Result

Figures 8 and 9 illustrate the performance of BABBLE across multiple audio and different tasks. BABBLE demonstrates strong performance in caption generation and fine-grained prediction tasks compared to the open-sourced audio models. Figure 9b shows an example where BABBLE struggles in understanding long audio.

## I Use of AI Assistant

We use the ChatGPT-5.1 model for occasional simple coding and polishing the manuscript. All text and results in this manuscript are originally produced by the authors.

**Algorithm 1** Generating Audio Mixtures

**Require:** primary source: home recordings $x_1(t)$, noise source: set of $N$ speech segments from secondary speaker sources $\{x_2^{(i)}(t)\}_{i=1}^{N}$, SNR values, $\mathrm{SNR}_{\mathrm{dB}} = 5dB$

**Ensure:** Mixed audio signal $y(t)$

**Compute power of primary source:**

$$P_1 = \frac{1}{T_1} \sum_{t=0}^{T_1} x_1^2(t) \tag{1}$$

**Compute the target signal for each noise source:**

**for** $i = 1$ to $N$ **do**

Compute power of each noise source:

$$P_2^{(i)} = \frac{1}{T_2^{(i)}} \sum_{t=0}^{T_2^{(i)}} \left( x_2^{(i)}(t) \right)^2 \tag{2}$$

Compute scaling factors to achieve target SNR:

$$\alpha_i = \sqrt{\frac{P_1}{P_2^{(i)} \cdot 10^{\mathrm{SNR}_{\mathrm{dB}}^{(i)}/10}}} \tag{3}$$

Scale the noise source:

$$x_2^{(i)\prime}(t) = \alpha_i \cdot x_2^{(i)}(t) \tag{4}$$

**end for**

**Randomly determine non-overlapping insertion times:**

Initialize empty list $\mathcal{T} = \emptyset$ for storing insertion intervals

**for** $i = 1$ to $N$ **do**

**repeat**

Sample $\tau_i \sim \mathcal{U}(0, T_1 - T_2^{(i)})$

**until** $\tau_i$ does not overlap with any interval in $\mathcal{T}$

Add $[\tau_i, \tau_i + T_2^{(i)}]$ to $\mathcal{T}$

**end for**

**Construct mixed audio signal:**

Initialize $y(t) = x_1(t)$

**for** $i = 1$ to $N$ **do**

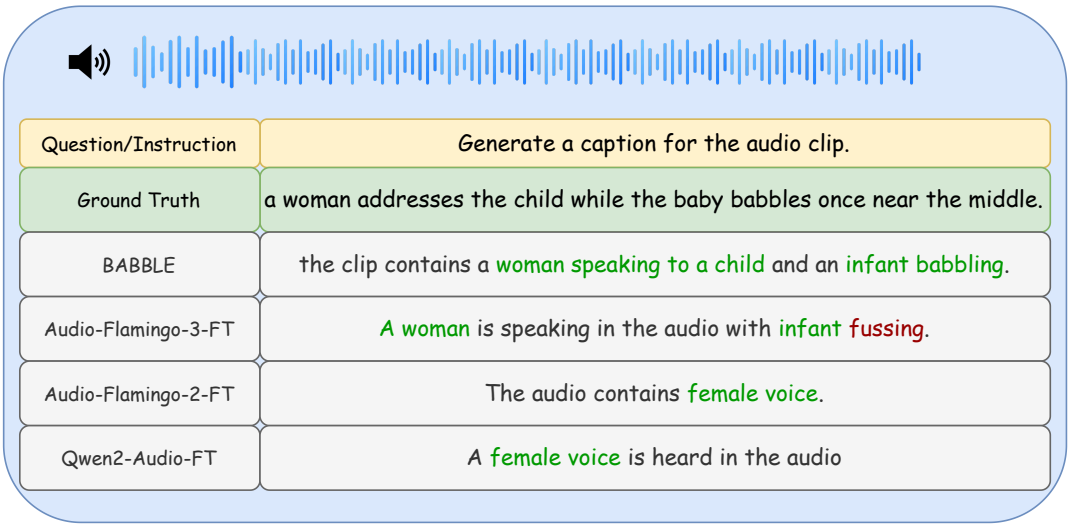
**for** $t \in [\tau_i, \tau_i + T_2^{(i)}]$ **do**

$$y(t) = y(t) + x_2^{(i)\prime}(t - \tau_i) \tag{5}$$

**end for**

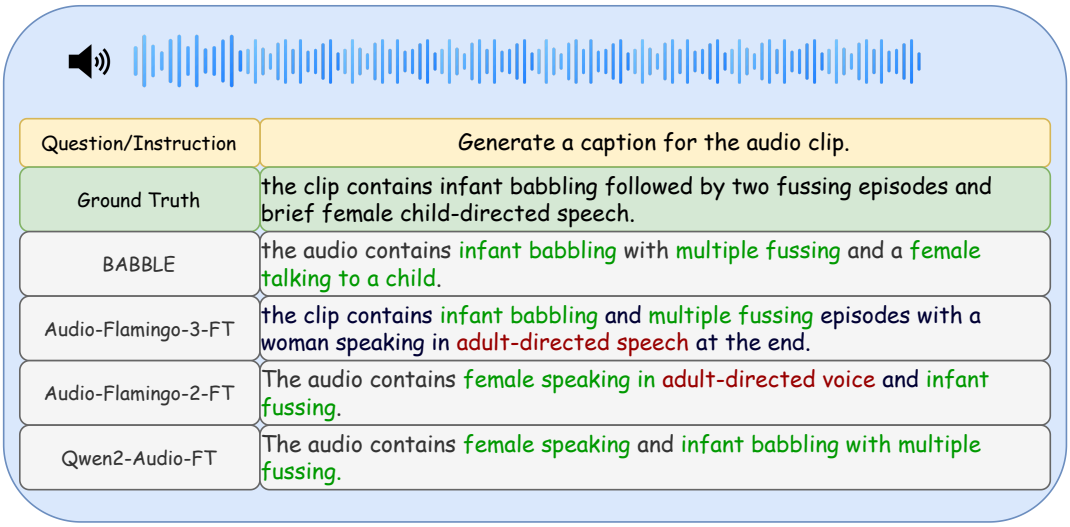
**end for**

**return** $y(t)$



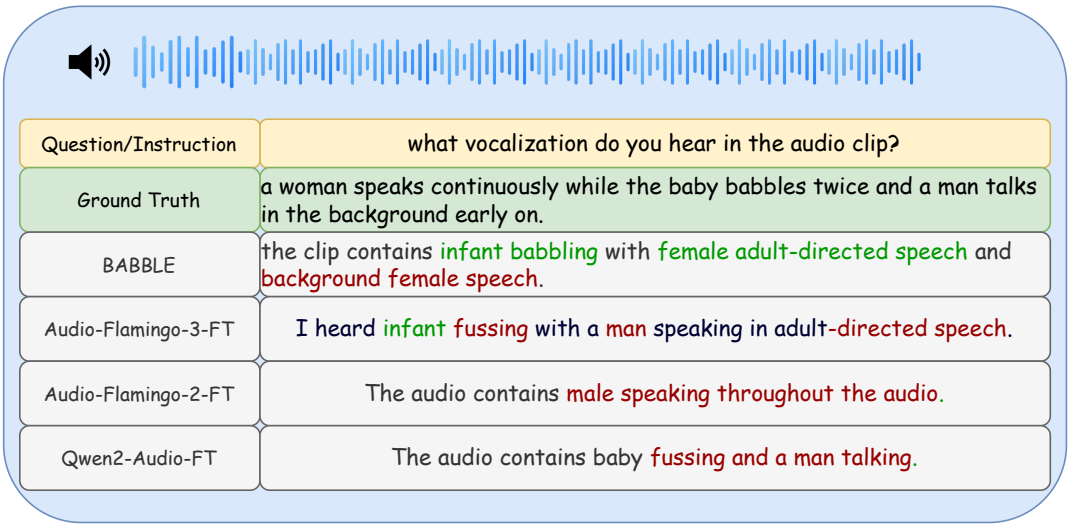
| Question/Instruction | Generate a caption for the audio clip. |
|---|---|
| Ground Truth | a woman addresses the child while the baby babbles once near the middle. |
| BABBLE | the clip contains a woman speaking to a child and an infant babbling. |
| Audio-Flamingo-3-FT | A woman is speaking in the audio with infant fussing. |
| Audio-Flamingo-2-FT | The audio contains female voice. |
| Qwen2-Audio-FT | A female voice is heard in the audio |

(a)



| Question/Instruction | Generate a caption for the audio clip. |
|---|---|
| Ground Truth | the clip contains infant babbling followed by two fussing episodes and brief female child-directed speech. |
| BABBLE | the audio contains infant babbling with multiple fussing and a female talking to a child. |
| Audio-Flamingo-3-FT | the clip contains infant babbling and multiple fussing episodes with a woman speaking in adult-directed speech at the end. |
| Audio-Flamingo-2-FT | The audio contains female speaking in adult-directed voice and infant fussing. |
| Qwen2-Audio-FT | The audio contains female speaking and infant babbling with multiple fussing. |

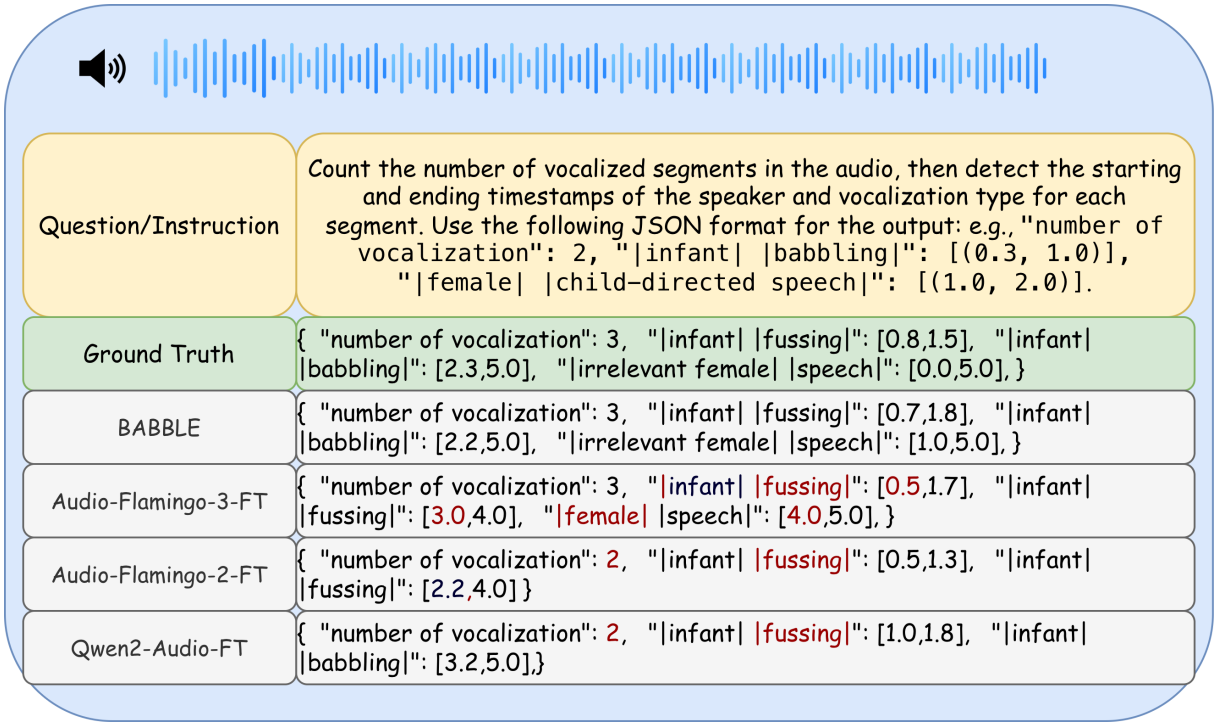
(b)



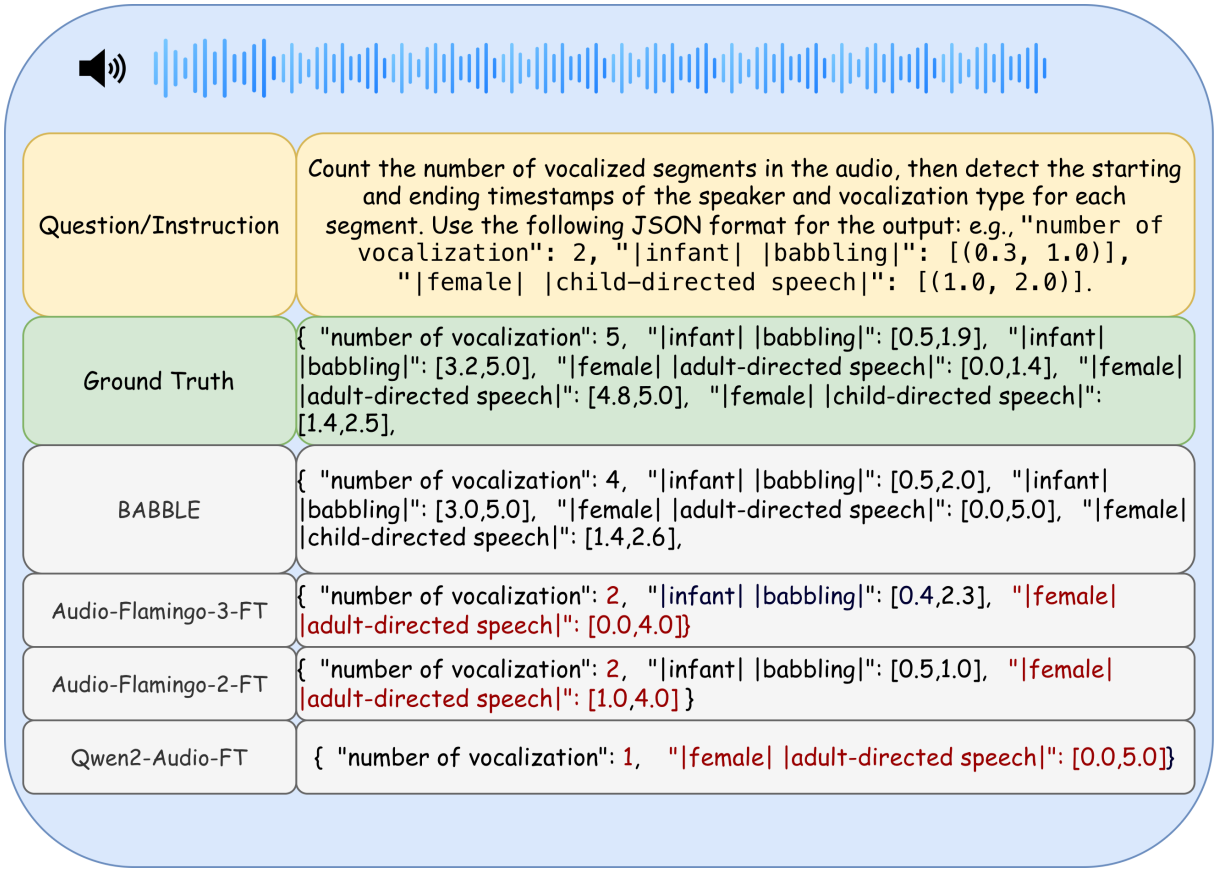
| Question/Instruction | what vocalization do you hear in the audio clip? |
|---|---|
| Ground Truth | a woman speaks continuously while the baby babbles twice and a man talks in the background early on. |
| BABBLE | the clip contains infant babbling with female adult-directed speech and background female speech. |
| Audio-Flamingo-3-FT | I heard infant fussing with a man speaking in adult-directed speech. |
| Audio-Flamingo-2-FT | The audio contains male speaking throughout the audio. |
| Qwen2-Audio-FT | The audio contains baby fussing and a man talking. |

(c)

Figure 8: Generated caption example from BABBLE, finetuned Audio-Flamingo-3, Audio-Flamingo-2, Qwen2-audio. While BABBLE consistently generates coherent captions with accurate speaker and vocalization types, other models struggle to get the correct predictions.




(a)




(b)

Figure 9: Qualitative results for event-based prediction from BABBLE and other fine-tuned audio-LLMs.