MoXaRt: Audio-Visual Object-Guided Sound Interaction for XR

Tianyu Xu
Google
Mountain View, CA, USA
tyx@google.com

Sieun Kim
University of Michigan
Ann Arbor, MI, USA
sieunk@umich.edu

Qianhui Zheng
University of Michigan
Ann Arbor, MI, USA
qianhui@umich.edu

Ruoyu Xu
Columbia University
New York, NY, USA
rx2241@columbia.edu

Tejasvi Ravi
Google
San Francisco, CA, USA
ravitejasvi@google.com

Anuva Kulkarni
Google
Mountain View, CA, USA
anuvak@google.com

Katrina Passarella-Ward
Google
San Francisco, CA, USA
kpassarella@google.com

Junyi Zhu*
University of Michigan
Ann Arbor, MI, USA
zhujunyi@umich.edu

Adarsh Kowdle*
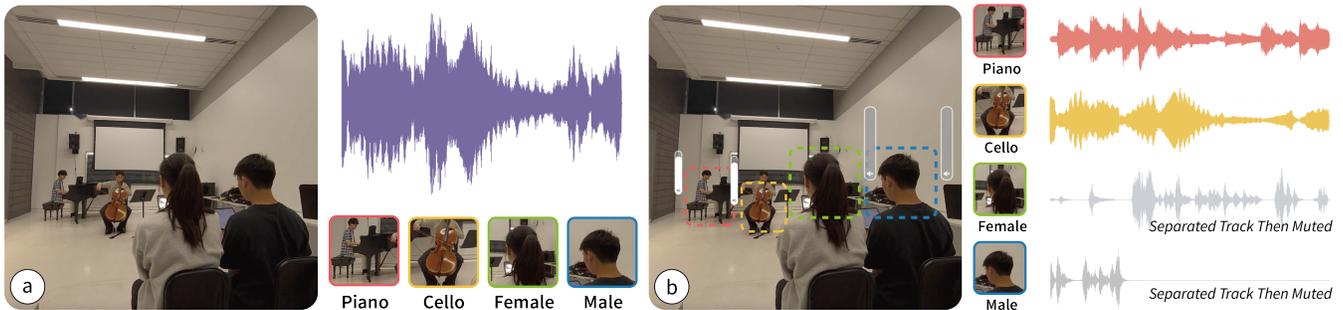Google
San Francisco, CA, USA
adarshkowdle@google.com

Figure 1: MoXaRt is a real-time XR system that (a) uses audio-visual cues to separate sound sources from a single audio channel and (b) enables fine-grained, interactive control over the user's soundscape.

Abstract

In Extended Reality (XR), complex acoustic environments often overwhelm users, compromising both scene awareness and social engagement due to entangled sound sources. We introduce MoXaRt, a real-time XR system that uses audio-visual cues to separate these sources and enable fine-grained sound interaction. MoXaRt's core is a cascaded architecture that performs coarse, audio-only separation in parallel with visual detection of sources (*e.g.,* faces, instruments). These visual anchors then guide refinement networks to isolate individual sources, separating complex mixes of up to 5 concurrent sources (*e.g.,* 2 voices + 3 instruments) with ~2 second latency. We validate MoXaRt through a technical evaluation on a new dataset of 30 one-minute recordings featuring concurrent speech and music, and a 22-participant user study. Empirical results

indicate that our system significantly enhances speech intelligibility, yielding a 36.2% ($p < 0.01$) increase in listening comprehension within adversarial acoustic environments while substantially reducing cognitive load ($p < 0.001$), thereby paving the way for more perceptive and socially adept XR experiences.

CCS Concepts

• **Human-centered computing** → **Mixed / augmented reality**.

Keywords

extended reality, sound separation, audio-visual, multimodal machine learning, large language models, vision language models

1 Introduction

Modern Extended Reality (XR) glasses and headsets grant users unprecedented control to filter, enhance, and overlay digital information onto their visual world [24, 37, 58]. Yet, this level of fine-grained, interactive control over perception has not extended to the auditory realm [32, 64]. In complex acoustic environments, XR users are often overwhelmed by a cacophony of sounds, making it difficult to focus on what truly matters [49]. This challenge stems from the user's fundamental inability to disentangle multiple simultaneous sound sources [8]. Unlike visual elements, which can be digitally manipulated, the ambient audio of people talking, instruments playing, and background noise remains a largely unfiltered and entangled mixture [64]. The result is compromised scene awareness and social engagement; important sounds, such as a specific conversation or a lead musician's solo, are drowned out, which impairs comprehension and increases cognitive load [49].

The core difficulty lies in performing real-time sound source separation on a head mounted device. The acoustic scene in XR is highly non-stationary, as head and sound source motion continually alter directions of arrival and room acoustics [17]. Furthermore, microphones on head-worn devices provide a limited spatial aperture due to their compact form factor, resulting in effectively mono or near-mono auditory inputs despite multiple sensors when sound sources are at distant (*e.g.,* > 3m) [38]. This hardware constraint, combined with the complexity of real-world auditory content—where human speech, music, and ambient noise constantly overlap and mask one another—makes robust audio separation based solely on a headset's microphone inputs an extremely challenging problem. Past approaches have either relied on cumbersome hardware setups *e.g.,* large microphone arrays unsuitable for wearable use [11], or have been computationally too expensive for real-time interaction [15, 22].

We introduce MoXaRt, a real-time XR system that tackles this challenge by leveraging audio-visual cues to separate sound sources from a single audio channel and enable fine-grained, interactive control over the user's soundscape. MoXaRt's core is a cascaded architecture operating in parallel: it performs a coarse, audio-only separation to split the soundscape into general categories (*e.g.,* speech, music, and noise) while concurrently running face and instrument detection to identify visual anchors in the video stream. These anchors then guide specialized refinement networks, which isolate individual speakers from the main speech track and specific instruments from the music track. This process generates a live audio-visual map of the environment, identifying "who or what is making each sound." Based on this map, MoXaRt empowers users with a personal "audio mixer" for the world, allowing them to isolate a single speaker in a noisy room, adjust the volume of individual instruments in a band, or remix their entire acoustic environment in real time.

We validate MoXaRt through both technical evaluation and a user study. For our technical evaluation, we collected a new, challenging real-world audio-visual dataset featuring complex mixtures of up to five concurrent speakers and three musical instruments. On this dataset, MoXaRt's sound separation performance surpasses state-of-the-art audio-visual models. Our cascaded architecture successfully separates up to four distinct human voices or a mix of five

total sources in real time, with a consistent processing latency of around 2 seconds. Finally, a 22-participant user study demonstrated MoXaRt's practical benefits across six real-world scenarios, from a bustling concert to a multi-group meeting. With MoXaRt enabled, participants performed significantly better at understanding target speech in noisy conditions, achieving an average improvement of 36.2% ($p = 0.0058$) in a multi-speaker listening comprehension task, and they reported significantly higher clarity and lower cognitive load.

In summary, we contribute:

- MoXaRt, the first XR system to integrate real-time, audio-visual sound separation as a core primitive for direct, interactive control over a user's authentic soundscape.
- A cascaded audio-visual transformer model that leverages visual cues to perform robust, real-time separation of multiple speech and non-speech sources from a single-channel audio input.
- A real-world audio-visual dataset featuring complex scenes with multiple concurrent speakers and musical instruments, designed for evaluating interactive sound separation systems.
- A comprehensive evaluation demonstrating our model's state-of-the-art technical performance and a user study confirming that MoXaRt significantly improves communication clarity, reduces cognitive load, and enhances user experience in complex auditory environments.

2 Related Work

2.1 Sound Interaction in XR

Audio is a crucial component for creating persuasive and immersive experiences in XR. Foundational research has centered on high-fidelity spatialization, using technologies like binaural rendering and Head-Related Transfer Functions (HRTFs) to create a sense of presence [12, 50]. Recent work continues to enhance this immersion by leveraging multimodal scene understanding to create context-aware auditory experiences [10, 33, 34, 48, 63], with industry leaders investing heavily in making virtual sounds perceptually indistinguishable from reality [31]. However, this blending of real and virtual soundscapes in Mixed Reality (MR) introduces new challenges. The accumulation of sounds from users' distinct physical environments often leads to acoustic incoherence and increased cognitive load [3, 30]. This necessitates a paradigm shift: moving from simply rendering sound to enabling users to manage and manipulate their auditory environment, a goal Haas et al. formalized as Interactive Auditory Mediated Reality (IAMR) [20].

In response to this need, the HCI community has developed compelling paradigms that treat audio as an active medium. A major thread focuses on augmenting or filtering the soundscape at a global or class level. For instance, Wu et al.'s *New Ears* allows users to spatially decouple their listening point from their physical location to facilitate object search [62], while Chang et al.'s *SoundShift* manages information overload by spatially shifting notifications or altering the transparency of real-world audio [3]. Similarly, semantic filtering systems like *Semantic Hearing* [55] and *ProtoSound* [25] enable users to suppress or prioritize specific sound categories (*e.g.,* vacuums, speech) or personalize sound events for accessibility.

More recently, research has shifted toward object-centric interaction. *SonoHaptics* maps visual properties to audio-haptic feedback to aid cursor-less gaze-based selection [9], while *AudioMiXR* treats virtual audio as a tangible object for 6DoF spatial manipulation [61]. Similarly, *SonifyAR* and *Sonify Anything* use visual recognition to synthesize context-aware sound effects [45, 48]. However, these systems rely on generative or synthetic audio rather than processing the object's actual acoustic emission. In the domain of veridical audio, *Look Once to Hear* enables users to isolate a specific speaker by looking at them [56]. Yet, this approach requires an explicit enrollment phase and functions as a binary filter (target vs. noise), lacking the ability to dynamically remix multiple concurrent sources (*e.g.,* music and speech). The critical limitation remains: no existing XR system allows users to seamlessly separate, balance, and remix the live, veridical sounds of multiple real-world objects in real-time.

2.2 Sound Separation in Machine Learning

Parallel to these HCI advancements, the signal processing and machine learning communities have developed sophisticated technologies for analyzing and separating sound sources. The classical cocktail party problem, which involves isolating individual speakers from a mixture of voices [8], has evolved from early signal processing methods like Independent Component Analysis (ICA) [23] and blind source separation (BSS) [2] to modern deep learning approaches that achieve remarkable separation quality [35].

2.2.1 Audio-Only and Universal Separation. Modern audio-only source separation is characterized by Transformer-based architectures, with models like MossFormer2 [66] achieving state-of-the-art results on speech separation benchmarks. In the music domain, specialized architectures like the Band Split Roformer (BSR) have achieved high-fidelity isolation of individual instruments [29, 42]. The field has also expanded to address the "cocktail fork problem" [40], separating complex soundtracks into speech, music, and sound effects simultaneously, while other work leverages spatial cues and linguistic context for simultaneous translation [7]. Universal sound separation systems [27] attempt to separate arbitrary sources without prior knowledge, though performance varies across categories. Building on these approaches, Target Sound Extraction (TSE) focuses on isolating specific sources based on text labels or reference audio [54, 57]. While recent work has explored real-time TSE [54] and wearable separation systems like *ClearBuds* and *Look Once to Hear* [4, 56], these typically function as binary extractors or speech enhancers, lacking the visual grounding and continuous, multi-source mixing control required for intuitive XR interaction.

2.2.2 Audio-Visual Separation. A significant breakthrough has been the incorporation of visual information to guide audio separation. This domain was pioneered by *The Sound of Pixels* [65] and *Looking to Listen* [16], which demonstrated that vision could act as a precise filter for both musical instruments and human speech. Building on this, *VisualVoice* introduced cross-modal consistency to leverage facial appearance as a strong prior for separating speech [18]. Modern architectures extend these principles: *AudioScopeV2* [53] uses cross-modal attention to bind on-screen objects to sounds, while *iQuery* formulates separation as a query-based task, utilizing detected visual objects to initialize learnable audio queries [5].

More recent work, like *Crab* [15] and *AVMossformer2* [67], employs transformer architectures to learn joint representations. Furthermore, advanced generative approaches, such as *DAVIS* [22], employ diffusion models to synthesize separated sound.

Despite these advances, current ML approaches face significant limitations when deployed for interactive applications. Most models are designed for offline processing, requiring the entire audio sequence for optimal performance [15, 16, 22, 53, 65]. Real-time operation, a prerequisite for interactive XR, often necessitates compromises in separation quality or introduces unacceptable latency and computational requirements that exceed the capabilities of consumer XR devices, particularly in generative diffusion models [22, 57]. Furthermore, these systems typically treat separation as a passive signal processing task without human input in the loop. For example, they can isolate a violin from an orchestra in a recording, but they cannot respond to a user pointing at a specific violin in a live performance and saying "make that one louder" while preserving the veridical sound of that instrument.

The preceding review highlights a critical disconnect between the high-level interaction needs of XR soundscapes and the low-level capabilities of current sound separation models. While HCI research manipulates live audio through global filters or generative synthesis, it lacks frameworks for object-centric, veridical interaction with live signals. Conversely, ML models isolate authentic sources but typically function as passive, offline tasks. Consequently, no prior work has operationalized visually-guided source separation as a real-time primitive for direct user interaction in XR.

3 MoXaRt

To enable intuitive control over complex acoustic scenes, we present MoXaRt, an XR system utilizing a cascaded, coarse-to-fine architecture for interactive audio-visual source separation. The user interface (Sec. 3.1) visualizes separated sources and allows users to adjust individual volumes for real-time audio remixing (Sec. 3.1.2). By offloading model inference to an external PC via wireless transmission (Fig. 2), the system ingests multi-modal streams through a parallel analysis stage (Sec. 3.2.1). The resulting outputs serve as contextual priors for specialized Music Refinement (Sec. 3.2.2) and Speech Refinement (Sec. 3.2.3) networks. The following sections detail MoXaRt's user interface and model architecture.

3.1 User Interface

3.1.1 User Interaction. Our headset-based XR interface visualizes localized sound sources, allowing users to control individual volumes. Each model-detected source is represented by a visual marker at its spatial position, enabling rapid identification of multiple concurrent sounds. Users can select specific sources and adjust their levels via interactive sliders (operated through controllers or hand gestures) to customize audio mixing for varying scenarios. To ensure an immersive experience, the remixed audio is delivered via noise-canceling earbuds connected to the XR headset, effectively isolating the user from ambient interference.

3.1.2 Implementation. We implemented MoXaRt for the Meta Quest 3 headset[1] using Unity 6000.2.2f1[2]. Model inference is

---

[1]Meta Quest 3. https://www.meta.com/quest/quest-3/
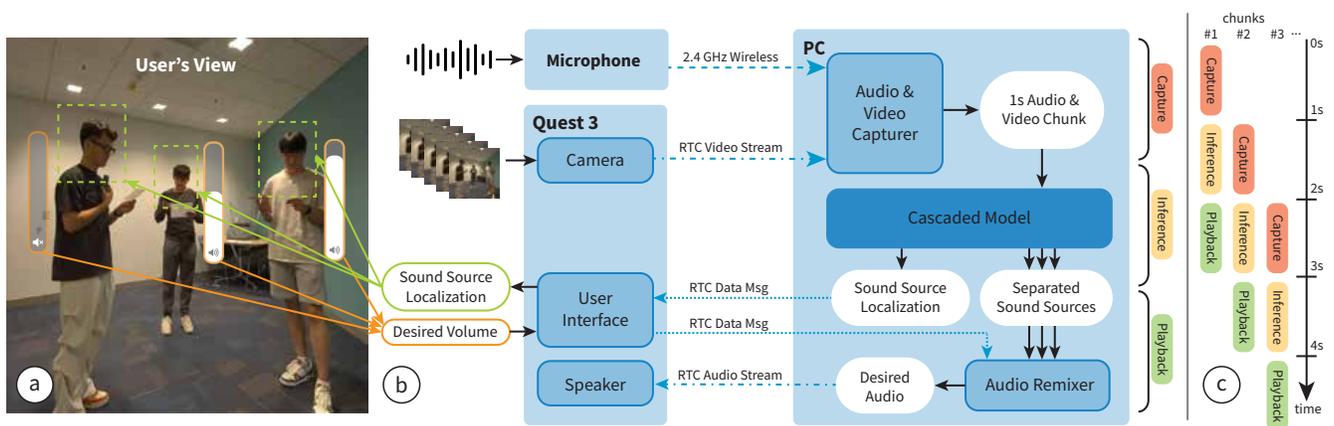[2]Unity. https://unity.com

Figure 2: The user interface of MoXaRt and walkthrough of its internal implementation. (a) The user can identify separated sound sources through visualizations and adjust the volume of each source for remixing. (b) Video captured from the Quest 3 headset is streamed to the PC for model inference, and the remixed audio is streamed back. (c) To ensure continuous playback of remixed audio, MoXaRt operates in a three-stage pipeline: Video and Audio Capture, Model Inference, and Playback.

offloaded to a dedicated PC equipped with an NVIDIA RTX 5080 GPU. To ensure high-quality audio capture, we use a Rode Wireless GO II[3] microphone, which transmits wirelessly to a receiver connected to the PC. The system utilizes WebRTC for bidirectional, real-time data transmission: the headset streams captured video to the PC, while the processed audio is returned to the headset. This remixed audio is played through active noise-canceling earbuds to eliminate acoustic interference between the processed output and the physical environment. We opted for an external microphone because Meta's audio API restricts access to raw signals, providing only heavily filtered audio. To maintain consistency with typical headset hardware, the external microphone is set to mono-mode for all applications and user study scenarios presented in this work.

Processing Pipeline. To support real-time sound separation and remixing, our interface operates in a three-stage pipeline: *Capture* (video and audio), *Inference* (model processing), and *Playback* (Fig. 2 (c)). Each stage processes data in 1-second segments, resulting in continuous playback with a constant latency of 2 seconds. During the *Capture* stage, camera frames from the Quest 3 and audio from the Rode microphone are synchronized into 1-second chunks. These chunks are appended to a 1-minute rolling buffer, which provides the necessary temporal context for our model to ensure consistent source separation. In the *Inference* stage, this 1-minute window serves as input to our cascaded model. The model identifies all speakers and musical instruments, separating them into individual tracks and providing spatial coordinates for localization. Only the final 1 second of the separated audio, corresponding to the most recent input chunk, is retained for remixing. Finally, in the *Playback* stage, the audio remixer applies user-defined volume levels to each source and merges them into a single output. This remixed stream is transmitted via WebRTC and played through the user's earbuds.

Sound Source Localization. Upon deployment in a new environment, MoXaRt processes the initial video stream through the cascaded model to identify and localize all individual sound sources.

³Rode Wireless GO II. https://rode.com/en-us/products/wirelessgoii

The model generates coordinates for each source within the video frames, which are transmitted to the headset via RTC data messages and visualized as markers in the interface (shown in green in Fig. 2 (a)). Users can then adjust the volume of specific sources using interactive bars (shown in orange in Fig. 2 (a)) via controllers or hand gestures. These updated volume levels are sent back to the PC via RTC data messages. This feedback loop ensures that separated audio from subsequent inference cycles is remixed according to user preferences before being returned to the headset through the established RTC audio stream.

Audio Remixing. The audio remixer module synthesizes the final auditory scene by combining the separated audio stems with the user-defined gain levels. Let the set of N refined audio stems from the preceding stages be $\{s'_1, s'_2, \ldots, s'_N\}$. Based on user input via the interface, a corresponding gain vector $g = [g_1, g_2, \ldots, g_N]$ is determined, where each $g_n \in [0, 10]$ represents the desired amplification factor for its respective source. The final monophonic audio output y_{final} is synthesized as a weighted sum:

$$y_{\text{final}} = \sum_{n=1}^{N} g_n s'_n$$

This composite signal is subsequently rendered into a binaural stereo stream to provide the user with a spatially coherent auditory experience.

3.2 Model

3.2.1 Coarse Sound Separation. The Coarse Sound Separation module provides an initial estimate of the constituent sources within the input audio. For this task, we adapt the separation module $\mathcal{M}^S$ from AudioScopeV2 [53] that is based on the dilated convolutional network with a learnable encoder-decoder architecture. The network, defined by the function f_θ, processes a raw mixed waveform $x \in \mathbb{R}^T$ to estimate $\mathcal{M} = 3$ source waveforms $\{\hat{s}_m\}_{m=1}^{\mathcal{M}}$, where $\hat{s}_m \in \mathbb{R}^T$ corresponding to speech, music, and environmental noise.
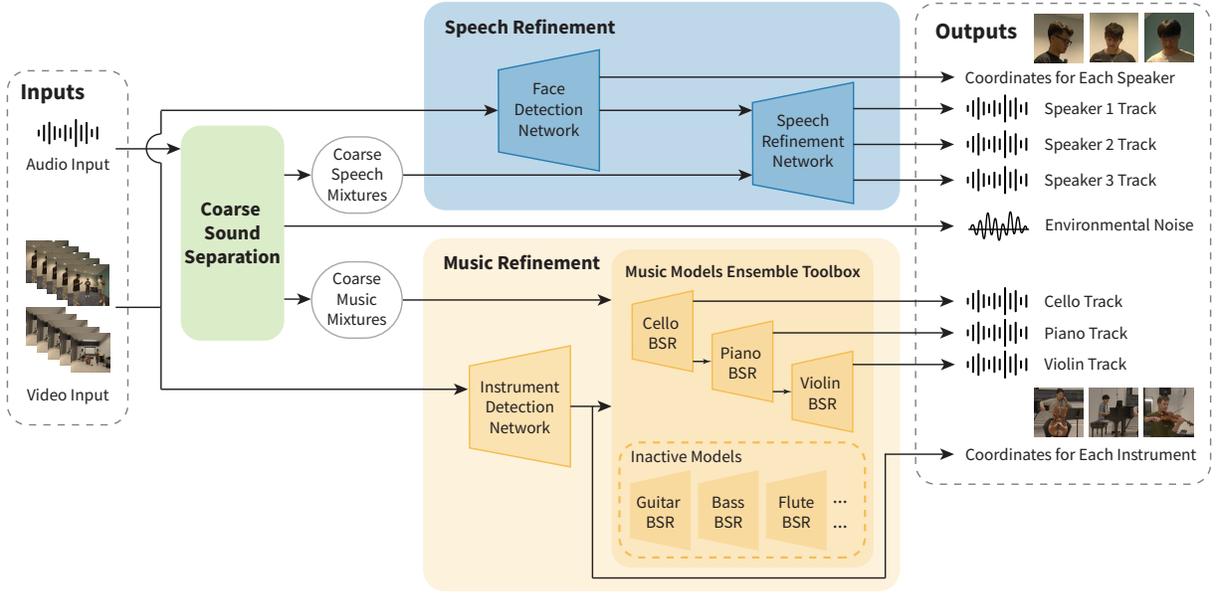
Figure 3: The cascaded architecture of MoXaRt for multi-modal sound separation. The system pipeline consists of two main stages: First, a Coarse Sound Separation module processes the raw audio inputs to produce initial coarse speech and music mixtures. Next, these mixtures, along with outputs from the Face Detection Network and Instrument Detection Network, are fed into two parallel refinement branches: Speech Refinement and Music Refinement. The Speech Refinement branch further disambiguates the coarse speech into individual speaker tracks and their corresponding spatial coordinates. The Music Refinement branch utilizes an ensemble of Band Split Roformer (BSR) models, dynamically activated by the detected instruments, to output isolated instrument tracks. This structured approach enables precise source separation by leveraging visual context only when specialized refinement is required.

The separation is performed by applying the model to the mixture:

$$\{\hat{s}_1, \hat{s}_2, \hat{s}_3\} = f_\theta(x)$$

Following the original training protocol, we incorporate a mixture consistency layer to enforce the constraint that the separated outputs sum to the original input:

$$x = \sum_{m=1}^{M} \hat{s}_m$$

In contrast to prior audio-visual methods [16, 22, 52], we adopt a purely audio-based approach for this initial stage. This design is motivated by two factors. First, it sidesteps the significant computational overhead of visual conditioning, which often provides only marginal gains in coarse separation. Furthermore, this modularity avoids the rigidity of monolithic end-to-end architectures; our pipeline enables dynamic resource allocation by activating specialized refinement modules only when their specific visual anchors are detected. Second, and more importantly for our cascaded design, it allows us to pre-train separation network on large-scale dataset corpora (*e.g.,* YFCC100M [51]) using self-supervised techniques like MixIT [59]. This robust, audio-only pre-training provides a generalized and high-quality initial separation, which is essential for the downstream Speech and Music Refinement modules to work effectively as illustrated in Fig. 3.

3.2.2 Music Refinement. To give users fine-grained control over musical elements, such as isolating an instrument for learning or creating a live karaoke mix, the Music Refinement module transforms the coarse music estimate into high-fidelity, separated instrument stems. It first employs a visual Instrument Detection network to identify which instruments are present. This visual context then dynamically activates specific models within our high-fidelity teacher model, whose knowledge is later distilled into an efficient student model.

Instrument Detection. First, a visual Instrument Detection network identifies which of $K = 7$ target instruments are present. For this, we use a multi-head DeepLabv3+ architecture [6], with a MobileNetV2 backbone [44], which is a standard choice for efficient on-device semantic segmentation. The backbone extracts multi-scale features which are processed by an Atrous Spatial Pyramid Pooling (ASPP) decoder. Following the DeepLabv3+ design, low-level features are fused with the ASPP output before being passed to a dedicated instrument classification head, which is trained to predict the presence of 7 relevant instrument classes: Vocal, Piano, Violin, Cello, Bass, Guitar, and Flute.

To train this model, we curated a large-scale dataset using an automated labeling pipeline. Starting with the COCO-Stuff dataset [1], we first employed the Segment Anything Model (SAM) [28] to generate high-quality instance masks for all objects in each image. Following the method of Peng et al. [39], a Vision-Language Model (VLM) then automatically assigned an instrument label to each

mask by comparing the semantic similarity of the cropped image region with text embeddings for our target classes. We trained the instrument classification head using a transfer learning strategy, freezing the pre-trained MobileNetV2 backbone and fine-tuning only the parameters of the new head and the ASPP decoder. We treat this as a multi-label classification problem and train the network using a Binary Cross-Entropy (BCE) loss with logits across the K instrument classes.

Dynamic Separation Ensemble (Teacher). Our high-fidelity teacher model for music is a dynamic ensemble of specialist networks. This module leverages predictions from the Instrument Detection network to perform targeted, high-quality separation. The ensemble consists of seven specialist models, one for each target instrument, based on the state-of-the-art Band Split Roformer (BSR) architecture [29]. To optimize efficiency, the system utilizes dynamic activation where only the BSR models corresponding to instruments detected with a confidence above a threshold ($\tau = 0.5$) are executed. This sparse and adaptive approach significantly reduces computational overhead while maintaining the high fidelity required to serve as an ideal data generator for our student model.

Distilled Real-Time Model (Student). Although the teacher ensemble provides state-of-the-art fidelity, its computational footprint is prohibitive for real-time XR applications. To address this, we employ model distillation [21]. Let the teacher be represented by a function T and the lightweight student model be S_ϕ with parameters ϕ. We generate a dataset of pairs consisting of the coarse track x_c and pseudo-target stems $\hat{y}_T$, where $\hat{y}_T = T(x_c)$. The student is trained to mimic the teacher by optimizing its parameters ϕ to minimize a signal-level loss $\mathcal{L}_{\text{signal}}$:

$$\min_\phi \mathcal{L}_{\text{signal}}(S_\phi(x_c), \hat{y}_T)$$

For $\mathcal{L}_{\text{signal}}$, we utilize the negative scale-invariant signal-to-distortion ratio (SI-SDR) loss [43]. This strategy effectively transfers the capabilities of the specialist ensemble into a single efficient model.

3.2.3 Speech Refinement. Concurrent with music refinement, this module resolves the coarse speech signal into discrete stems for each speaker. The system utilizes the video feed to disambiguate speakers based on their facial information, allowing the audio to be mapped to specific individuals in the scene.

Face Detection. To associate speech with a specific speaker, we adapt the yolov8-face model [14]. This open-source project fine-tunes the YOLOv8 framework [26] specifically for face detection tasks. The model provides real-time bounding boxes that serve as the visual anchors required to guide the downstream separation process.

Online and Offline Separation Models. Our system supports two distinct processing modes. For offline processing, we employ AV-MossFormer2 [67] as a high-fidelity teacher model, T_{AV}. For online processing, we distill the teacher's knowledge into a lightweight student model, S_ψ, which is based on the AudioScopeV2 [53] architecture. Given a coarse speech track x_{sp} and the corresponding visual face crops v, the student is trained on the teacher outputs

$\hat{y}_{T_{AV}} = T_{AV}(x_{sp}, v)$. The student parameters ψ are optimized by minimizing the same distillation loss described in Sec. 3.2.2.

3.2.4 Implementation. We deployed and evaluated MoXaRt using the hardware setup described in Sec. 3.1.2. The visual frontend consists of the face detection model (3.2M parameters, 12.1 MB) and the instrument detection model (33M parameters, 132 MB). The audio pipeline begins with the Coarse Sound Separation module (6M parameters, 24 MB), followed by the distilled online student models for Speech Refinement (9M parameters, 36 MB) and Music Refinement (7M parameters, 28 MB). In the current configuration using these student models, the end-to-end processing latency is approximately 2 s. While these models are designed to operate within an interactive budget, this latency figure serves as a functional baseline for the cascaded architecture. We provide a comprehensive breakdown of latency components and discuss paths toward sub-second real-time performance in Sec. 7.1.

4 Application Scenarios

The ability of MoXaRt to dissect a complex acoustic scene into its constituent parts and grant the user interactive control over them unlocks a wide range of applications. We highlight three scenarios that demonstrate the system's potential to reshape how we experience and interact with our auditory environment.
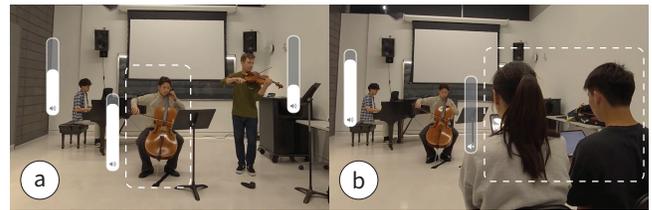

Figure 4: Interactive Music Experiences. MoXaRt enables users to actively control their listening experience during live music events. (a) The user can act as a real-time audio engineer, selectively adjusting the gain of individual instruments, such as the violin, cello, and piano in a trio—to create a personalized mix. (b) In a concert setting, the user can enhance acoustic focus by isolating the musicians' performance from distracting audience chatter.

4.1 Interactive Music Experiences

In settings like live concerts or music rehearsals, MoXaRt transforms the user from a passive listener into an active participant in the audio experience (Fig. 4). A user attending a concert can act as their own real-time sound engineer, adjusting the audio mix to their personal preference—for instance, boosting the violin during a solo or lowering the piano to focus on the cello's bassline. This capability is also valuable for music education, as a student could isolate a specific instrument within an ensemble to better study its part. Furthermore, the system enhances listening clarity by functioning as an intelligent de-noiser. By isolating the desired performance, a user can suppress distracting ambient sounds, such as audience chatter, to maintain an immersive and focused experience.
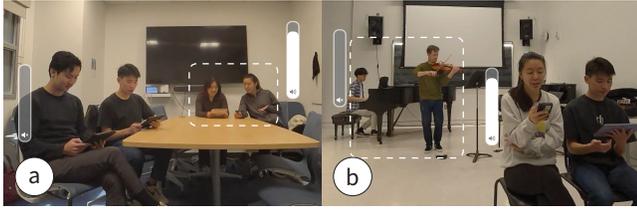
4.2 Enhanced Social Interaction

Figure 5: Enhanced Social Interaction. MoXaRt facilitates communication in complex and noisy social settings. (a) In an environment with multiple simultaneous conversations, a user can selectively amplify a target conversation to maintain focus. (b) The system can improve speech intelligibility in chaotic environments by isolating a conversation partner's voice from loud background noise, such as live music.

MoXaRt provides a powerful tool for navigating the challenges of complex social environments, effectively offering a practical solution to the classic "cocktail party problem" (Fig. 5). At a crowded reception or in a noisy café, a user can dynamically manage their auditory focus and stay engaged in multiple conversations. For example, they can choose to amplify a specific conversation, even one happening at a distance, while attenuating other nearby voices and background noise. This selective attention reduces the cognitive load associated with straining to hear in loud settings, making social engagement less fatiguing and more accessible, particularly for individuals with hearing difficulties. The user can seamlessly switch their focus between different conversational groups, giving them unprecedented control over their social soundscape.

4.3 Future Application: Context-Aware AI Assistance

Beyond direct human consumption, the clean, object-centric audio-visual streams produced by MoXaRt serve as a critical primitive for intelligent, context-aware AI assistants[19, 36]. Current voice-based agents typically operate on a single, mixed-audio stream, causing them to fail in "cocktail party" scenarios where overlapping speech confounds Automatic Speech Recognition (ASR) engines.

By feeding MoXaRt's separated audio outputs into downstream models, we can unlock two specific technical capabilities that are currently intractable in mixed-reality environments:

4.3.1 Real-Time Diarized Translation. In a multi-lingual meeting, MoXaRt can enable per-speaker translation. The system buffers the separated audio stream s_i' for each detected speaker. These streams are processed in parallel by a streaming ASR model to generate text. Because MoXaRt reduces the Word Error Rate (WER) by significant margins compared to raw audio (as shown in Table 2), the ASR engine receives a signal that is actually intelligible, unlike the raw mixture. The resulting text is translated and rendered as a subtitle floating above the corresponding speaker's coordinate box (Fig. 6). The primary challenge here is cascading latency. The total latency $L_{total} = L_{sep} + L_{ASR} + L_{trans}$. While MoXaRt 's current

separation latency is $\approx 2s$, optimizing the pipeline to use sliding-window inference with smaller chunks could bring this closer to the perceptual threshold required for live captioning.

4.3.2 Queryable RAG-based Audio-Visual Memory. Current memory in AI assistants is limited because it cannot robustly distinguish who said what in a crowded room. MoXaRt enables a visually-grounded Retrieval-Augmented Generation (RAG) pipeline. Instead of discarding the separated audio buffers, the system can transcribe them into speaker-tagged text segments. These segments are embedded and stored in a vector database. A user can then query an LLM: "What did Peter say about the deadline?". The LLM retrieves the specific speech segment associated with Peter's visual ID. This application relies on the system's ability to maintain a persistent identity for sound sources over time. As noted in Sec. 7.3, future iterations could leverage stronger visual identification embeddings to ensure that the Speaker tag remains consistent throughout long-duration interactions.
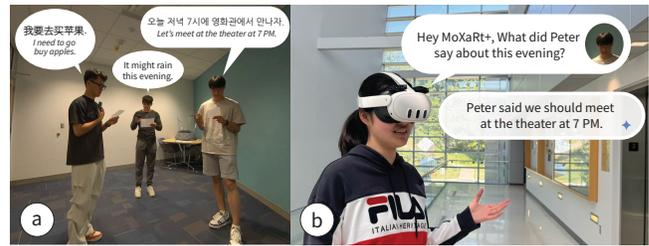

Figure 6: Downstream AI Assistance. The separated streams from MoXaRt can serve as clean, audio-visually aligned input for LLM assistance (*e.g.,* Gemini, Chatgpt.) (a) The system can isolate individual speakers in a multi-lingual environment, enabling real-time, multi-person transcription and translation. (b) The system's output could be used to create a queryable log of interactions, allowing a user to ask questions about past events (*e.g.,* "What did Peter say about this evening?") and receive a contextually grounded response.

5 Technical Evaluation

We performed a technical evaluation to assess the performance of MoXaRt. Our goal was to quantitatively measure the audio separation quality of our system and validate the separation performance of our distilled real-time models. To do this, we compared MoXaRt against several state-of-the-art baselines using objective metrics for both speech and music separation tasks. A supplemental evaluation on the trade-off between model performance and interactive latency is provided in Sec. 7.1. This section details our experimental setup and presents an analysis of the results.

5.1 Methods

5.1.1 Dataset. To evaluate MoXaRt in realistic and challenging acoustic environments, we collected a new audio-visual dataset, as existing corpora do not adequately capture the complex, mixed-source scenarios relevant to interactive XR applications. Our dataset consists of 30 unique, one-minute-long recordings categorized into three main types: speech-only, music-only, and mixed-source.

Table 1: Comparison of MoXaRt against baselines and audio-visual sound separation benchmarks.

Reference System	Architecture	Target Domain	Object Selection	Max Sources	Real-Time
Sound of Pixels	End-to-End	Music Only	✓	2	✓
DAVIS	End-to-End	Universal†	✗	3	✗
AudioScopeV2	Cascaded	Universal	✗	4	✓
AV-MossFormer2	End-to-End	Speech Only	✗	3	✗
MoXaRt (Offline)	**Cascaded**	**Universal**	✓	**8**	✗
MoXaRt (Real-Time)	**Cascaded**	**Universal**	✓	**8**	✓

Note: Universal denotes models capable of separating both speech and music sources. Universal† indicates the architecture is domain-agnostic, but requires separate training for specific domains (speech vs. music).

The speech-only subset contains 15 recordings designed to test performance on multi-speaker conversations of increasing complexity. These scenarios feature one (C1), two (C2), three (C3), four (C4), and five (C5) concurrent speakers, with three unique recordings for each condition. The speaker configurations simulate realistic social gatherings, ranging from individuals speaking to multiple simultaneous conversations. For instance, the five-speaker condition features two concurrent conversational pairs and one additional independent speaker. The music-only subset comprises 9 recordings of instrumental performances featuring guitar, flute, piano, and violin. These include solo performances (C6), various duet combinations (C7), and trios (C8). Finally, the mixed-source subset contains 6 recordings that combine instrumental music and speech, featuring scenarios with one speaker accompanied by two instruments (C9) and two speakers accompanied by a single instrument (C10).

We recorded this dataset from a user's egocentric viewpoint. Our setup (Fig. 10) featured a Meta Quest 3 headset on a stand, outfitted with a Rode Wireless GO II wireless system to capture mono audio input. A tripod-mounted iPhone recorded a corresponding third-person video of each scene. Further details on the recording environment and procedure are located in Appendix A.

Figure 7: Data collection setup. (a) A Meta Quest 3 with two attached Rode Wireless GO II microphones captures the input mono-audio, while a stationary iPhone records a third-person view of the scene. (b) The three sound sources (two instruments and one speaker), each close-mic'd to record their individual ground-truth audio signal.

5.1.2 Baselines. To contextualize MoXaRt's performance, we compared it against a spectrum of baselines representing diverse and challenging architectural paradigms. Table 1 provides a high-level comparison of our system's architectural capabilities against these baselines. We selected the following systems for comparison:

- Raw Audio: We use the original, unprocessed audio to establish a lower performance bound for all metrics.
- Sound of Pixels [65]: An end-to-end pixel-level analysis method for filtering specific sound sources, primarily designed for musical instruments.
- AudioScopeV2 [53]: An on-screen sound separation system that employs cross-modal and self-attention networks to capture fine-grained audio-visual dependencies. Its inclusion is particularly relevant as our online student models are derived from its architecture.
- AV-Mossformer2 [66]: We use the AV MossFormer2 TSE 16K model from the ClearerVoice-Studio project [67]. This system serves as a strong point of comparison for contemporary audio-visual methods.
- DAVIS [22]: A state-of-the-art generative baseline for audio-visual separation that employs diffusion and flow matching. To ensure a fair comparison, we trained a 3-source flow matching model (num_mix=3) using the official recipe. See Appendix B for training details.

5.1.3 Metrics. To quantitatively evaluate the performance of MoXaRt and the baselines, we use two widely-adopted sets of metrics:

- Word Error Rate (WER): To measure speech intelligibility, we use Gemini 2.5 Pro to transcribe the separated speech audio and calculate the WER against the ground truth transcript.
- Deep Noise Suppression Mean Opinion Score (DNSMOS): To assess perceptual audio quality, we use the DNSMOS [41] metric (scale 1–5). We report the three standard submetrics: SIG (signal clarity), BAK (background suppression), and OVRL (overall quality).

5.1.4 Results and Discussion. The results of our objective evaluation on separation accuracy and quality are presented in Table 2. The main takeaways are as follows:

(1) MoXaRt Offline Achieves Superior Intelligibility. Our offline model, **MoXaRt Offline**, achieves the best performance in terms of speech intelligibility, securing the lowest **Word Error Rate (WER)** of **0.3824**. This result surpasses all baselines, including the highly competitive AV-Mossformer2.

(2) Trade-offs in Perceptual Quality. While MoXaRt excels in intelligibility, **AV-Mossformer2** obtains higher scores on the **DNSMOS** metrics. This suggests it produces audio with

Table 2: Objective evaluation of separation accuracy and quality.

Method	WER[†] ↓	DNSMOS-SIG ↑	DNSMOS-BAK ↑	DNSMOS-OVRL ↑
Raw Audio	0.6297 (± 0.4371)	3.7740 (± 0.1977)	2.4796 (± 0.3928)	2.8944 (± 0.1673)
Sound of Pixels	0.8636 (± 0.1787)	3.3212 (± 0.2467)	2.9535 (± 0.2935)	2.7291 (± 0.2004)
DAVIS	0.7259 (± 0.2484)	3.3212 (± 0.2467)	2.9535 (± 0.2935)	2.7291 (± 0.2004)
AudioScopeV2	0.5263 (± 0.3358)	3.5027 (± 0.5029)	3.0236 (± 0.4497)	2.9349 (± 0.3074)
AV-Mossformer2[†]	0.3956 (± 0.2847)	3.7245 (± 0.2813)	3.2564 (± 0.4728)	3.0759 (± 0.2892)
MoXaRt Offline	**0.3824 (± 0.2775)**	**3.6625 (± 0.3059)**	**3.019 (± 0.6078)**	**2.9949 (± 0.2585)**
MoXaRt Real-Time	**0.4990 (± 0.3156)**	**3.5015 (± 0.4976)**	**3.0065 (± 0.4266)**	**2.9364 (± 0.3060)**

[†] Calculations are done over the sessions that contain human speech.

slightly better perceived signal clarity and background suppression, potentially at the cost of some word-level accuracy.

(3) Real-Time Model Validates Distillation. The **MoXaRt Real-Time** model demonstrates the effectiveness of our knowledge distillation approach. It significantly outperforms its architectural basis, **AudioScopeV2**, achieving a much lower **WER (0.4990 vs. 0.5263)** while maintaining nearly identical DNSMOS scores.

(4) Performance of Other Baselines. The generative baseline, **DAVIS**, performed poorly on our dataset with its default configuration, likely due to sensitivity to acoustic characteristics without specific hyperparameter tuning. **Sound of Pixels** similarly struggled with the complex mixtures in our dataset, resulting in higher WER compared to our cascaded approach.

6 User Study

We evaluated MoXaRt's performance and user experience through a within-subjects study (N=22) across six challenging XR scenarios. Our mixed-methods approach contains both objective performance on listening comprehension tasks and subjective 1-10 scale ratings of key experiential factors.

6.1 Participants

We recruited 22 participants (8 female, 14 male; mean age 23.2, STD = 2.72). The group included 5 native and 17 non-native English speakers to provide a diverse cohort for the communication tasks. All participants reported normal hearing capabilities and were compensated $30 for their time.

6.2 Apparatus and Environment

The study was conducted using a Meta Quest 3 headset. To create dynamic and realistic yet controlled environments, scenarios involving conversations and musical performances featured pre-recorded actors and professional musicians via Quest's see-through camera in panoramic view. The actors performed standardized TOEFL scripts, while the instrumentalists (a piano, violin, and cello trio) performed both synchronized and unsynchronized pieces. The professional musicians and actors involved in the setup were compensated $100

for their contribution. We chose to record these scenarios and process them using a consistent real-time model to control for variance in speech and musical performance across participants.

6.3 Tasks and Scenarios

Each participant completed all six scenarios both with and without MoXaRt (baseline), presented in a counterbalanced order.

6.3.1 Task 1: Listening Comprehension. Participants were instructed to focus on a target conversation and subsequently answer TOEFL comprehension questions. The task was conducted across three distinct scenarios, with one easy, one medium and one hard in difficulty level from Educational Testing Service (ETS) TOEFL Practice Online (TPO) tests[4]. Two conversations of equal difficulty were used for each scenario, and both the order of conditions (MoXaRt and Baseline) and the pairing of conversations with conditions were counterbalanced across participants:

- **Scenario 1: Social Dining with Conversation over Music.** Focusing on a companion's voice over loud background music from a piano bar. The conversations lasted 2 minutes 50 seconds and 3 minutes respectively.
- **Scenario 2: Multi-Speaker Conference with Distant Male Voices.** Focusing on a more distant conversation between male speakers while a competing conversation occurred nearby. Each conversation lasted 2 minutes 45 seconds.
- **Scenario 3: Multi-Speaker Conference with Distant Female Voices.** Focusing on a more distant conversation between female speakers while a competing conversation occurred nearby. Each conversation lasted 3 minutes 30 seconds.

6.3.2 Task 2: Interactive Audio Remixing. Participants could freely manipulate the volume of individual sound sources in real-time. This task was conducted in three scenarios:

- **Scenario 4: Orchestra Evaluation.** Adopting a music teacher's perspective to isolate and evaluate one student's performance within a synchronized trio. Each performance lasted 3 minutes 10 seconds.

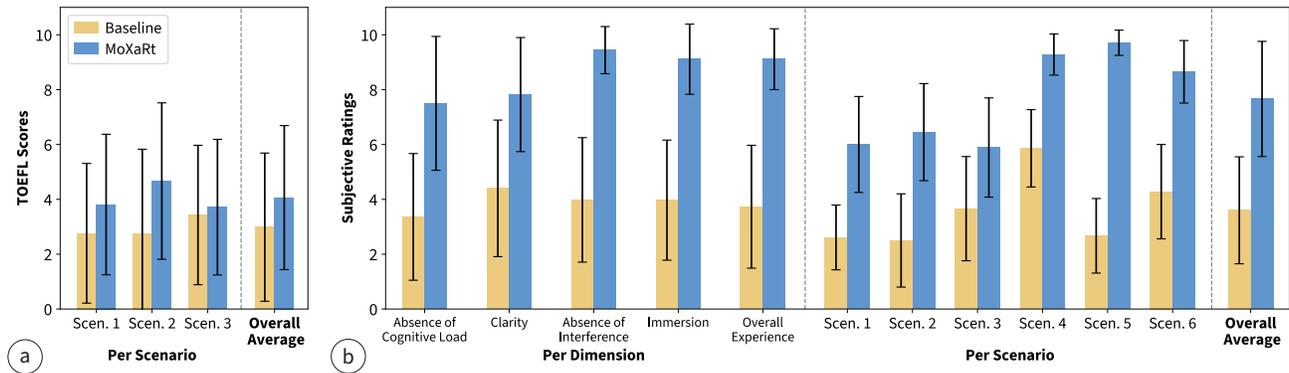---

[4]https://www.ets.org/toefl.html

Figure 8: User study results. (a) Average Listening Comprehension Performance per scenario and the overall average. MoXaRt led to a statistically significant 36.2% improvement in scores compared to the baseline (p = 0.0058). (b) Average subjective ratings (on a 10-point scale, higher scores indicate better experience) per dimension for the five core experiential dimensions and per scenario and the overall average. Participants rated their experience significantly higher with MoXaRt compared to the baseline (all p < 0.001). Error bars on the average represent the standard deviation.

- **Scenario 5: Rehearsal Evaluation.** Adopting the same teacher perspective, but in a chaotic environment where the trio's members practiced unsynchronized pieces. Each performance lasted 3 minutes 9 seconds.
- **Scenario 6: Concert Interference.** Adopting an audience member's perspective to filter out distracting chatter from nearby people and focus on the musical performance. The performances lasted 2 minutes 53 seconds and 3 minutes 1 second respectively.

For scenarios 4 and 5, the specific instruments participants evaluated were pre-assigned and counterbalanced.

6.4 Procedure

Upon arrival, participants were briefed on the study's purpose and gave informed consent. They were then familiarized with the XR equipment and the interaction mechanics of MoXaRt during a 5-minute tutorial session. Before the main tasks, each participant completed a practice trial to ensure they understood the objectives.

Each scenario was completed under both the MoXaRt Enabled and Baseline conditions. To counterbalance the condition order, the conversation-to-condition pairings, and the target instrument designations (piano, violin, or cello), the study was designed for multiples of 12 participants (2×2×3). Although 24 participants were originally recruited, two were unable to attend. As a result, the first 12 participants were fully counterbalanced, while the remaining 10 followed a randomized order.

Prior to each scenario, participants were given 30 seconds to preview the questionnaire. This allowed them to familiarize themselves with the specific conversation details to monitor, the target musical instrument for evaluation, and the criteria for subjective rating. After each condition within a scenario, they removed the headset to answer the relevant TOEFL questions within 3 minutes and complete the subjective rating questionnaire within 1 minute. A short break was provided between each scenario. The entire study session lasted approximately 60 minutes.

6.5 Measures

We collected both objective and subjective data to evaluate the system's performance and the user's experience.

6.5.1 Objective Measures. Our primary objective metric was Listening Comprehension Performance. This was measured by the participant's score on the TOEFL questions following each scenario, with 2 points awarded for each correct answer. A higher score indicates a greater ability to successfully comprehend and retain information amidst distractions.

6.5.2 Subjective Measures. After each scenario, participants provided subjective ratings on a 1-10 scale to quantify their experience. These ratings captured five core dimensions: Absence of Cognitive Load, Clarity, Absence of Interference, Immersion, and Overall Experience.

6.6 Results

6.6.1 Objective Results. To objectively evaluate the impact of MoXaRt on a user's ability to understand speech in noisy environments, we analyzed the Listening Comprehension Performance scores from Scenarios 1, 2, and 3. For Scenario 1, participants achieved a 38% higher score with MoXaRt (M=3.81, SD=2.56) compared to the baseline condition (M=2.76, SD=2.55). For Scenario 2, participants achieved a 69% higher score with MoXaRt (M=4.67, SD=2.85) compared to the baseline condition (M=2.76, SD=3.06). For Scenario 3, participants achieved a 8% higher score with MoXaRt (M=3.71, SD=2.47) compared to the baseline condition (M=3.43, SD=2.54). A two-way RM ANOVA was used to analyze the performance scores and Aligned Rank Transform [60] was applied before ANOVA to correct non-normality. ANOVA revealed that there were significant effects of *condition* (F = 11.955, p = 0.00248) to the performance, while *scenario* had no significant effect (F = 0.402, p = 0.672). There was no significant *condition × scenario* interaction effect of performance (F = 1.343, p = 0.272).

In addition, participants on average achieved a 36.2% higher total score (out of 10 points) with MoXaRt (M=4.06, SD=2.06) compared

to the baseline condition (M=2.98, SD=2.05). This improvement in performance was statistically significant (p = 0.0058), underscoring the substantial and reliable benefit of using the system. Notice that 1 out of 22 participants (P9) is excluded from the objective result analysis, since their result is >2.5 SDs outside of the scoring distribution of all participants, therefore an outlier.

These results are summarized in Fig. 8 (a). The data reveals a consistent trend where participants achieved significantly higher scores with MoXaRt enabled compared to the baseline. Across all three tested scenarios, our system yielded superior performance metrics. Notably, the most substantial improvements were observed in participants who performed poorly in the baseline condition; these individuals showed drastic performance gains when using our system, demonstrating its effectiveness in acoustic environments that users find particularly challenging.

6.6.2 Subjective Results. To evaluate user experience, we analyzed 1-10 scale ratings from the interactive audio tasks. Paired-samples t-tests were conducted to compare the MoXaRt Enabled and Baseline conditions across five core dimensions. Participants rated their experience significantly higher with MoXaRt across every dimension ($p < 0.001$). Specifically, they reported a significantly greater absence of cognitive load ($M = 7.50, SD = 2.44$ vs. $M = 3.36, SD = 2.31$), higher Clarity ($M = 7.82, SD = 2.08$ vs. $M = 4.40, SD = 2.49$), reduced interference ($M = 9.44, SD = 0.86$ vs. $M = 3.98, SD = 2.27$), and greater immersion ($M = 9.11, SD = 1.28$ vs. $M = 3.97, SD = 2.19$). Reflecting these benefits, the Overall Experience with MoXaRt ($M = 9.11, SD = 1.11$) was rated significantly higher than the baseline ($M = 3.73, SD = 2.24$).

We further analyzed subjective ratings across individual scenarios. Consistent with the dimensional results, participants rated their experience significantly higher with MoXaRt in every scenario ($p < 0.001$). Performance gains were substantial across the board: improvements over the baseline ranged from a 58% increase in Scenario 4 ($M = 9.28$ vs. $M = 5.86$) to a 264% increase in Scenario 5 ($M = 9.71$ vs. $M = 2.67$). Full details for all six scenarios, including Scenarios 1, 2, 3, and 6 (which saw increases of 130%, 158%, 61%, and 102% respectively), are visualized in Fig. 8 (b).

On average, participants provided 113% higher ratings with MoXaRt ($M = 7.66, SD = 2.10$) compared to the baseline ($M = 3.60, SD = 1.95$), indicating a robust subjective preference ($p < 0.001$). These findings complement the objective results, demonstrating that performance gains were accompanied by a substantially improved and less mentally demanding user experience.

7 Discussion and Future Work

Our work demonstrates the feasibility of utilizing real-time, visually-guided source separation as a core primitive for interactive XR audio experiences. While our technical and user evaluations validate the efficacy of the proposed approach, they also reveal specific limitations that provide a roadmap for future research.

7.1 Model Performance vs. Real-time Analysis

The overall user experience in MoXaRt depends on the balance between two competing factors: audio separation accuracy and interactive latency. Providing models with longer context windows

generally yields a superior listening experience with fewer artifacts (*e.g.*, lower WER), thereby increasing perceptual clarity at the expense of longer system response times. Conversely, aggressive parallel streaming reduces delay and enhances responsiveness to user actions, but risks degraded separation quality and increased computational overhead.

In MoXaRt, these factors are primarily controlled by the *window_size* and *chunk_size* parameters. The *window_size* determines the duration of the temporal context provided to the model, impacting both separation accuracy and inference time. The *chunk_size* dictates the frequency of model execution, which governs the computational load and audio playback latency. In our current implementation, we set these parameters to 60 seconds and 1 second, respectively (see Sec. 3.1.2). To quantify these trade-offs, we conducted experiments evaluating the effects of *window_size* and *chunk_size* on inference time, WER, computational burden, and total interactive latency.

To examine the trade-off between audio separation accuracy and interactive latency, we measured WER and model inference time across a range of window sizes on a log-scale up to 60 seconds. We processed our dataset at each window size and concatenated the results to calculate the corresponding WER. As illustrated in Fig. 9 (a), the highest separation accuracy (lowest WER) is achieved with the largest *window_size*, albeit at the cost of the longest inference time.

To evaluate total system delay, we measured the interactive latency using the smallest operational *chunk_size* for each given *window_size*. Although *chunk_size* can theoretically be reduced indefinitely, it is lower-bounded by the computational overhead required to maintain continuous streaming. We define interactive latency as the delay between a visual event and its corresponding audio rendering through MoXaRt, averaged over three measurements across three experimental runs. As shown in Fig. 9 (b), interactive latency decreases alongside *window_size*. By utilizing a 0.15-second chunk size with a 1-second window, we achieved an interactive latency of 0.82 seconds. These results suggest that chunk and window sizes can be jointly optimized to balance system responsiveness with separation quality. Further investigation is required to determine how such delays are perceived by users and their ultimate impact on the subjective experience.

7.2 Deployment Constraints

While our results show a promising path towards interactive audio-visual XR, the current implementation faces practical deployment constraints. First, the reliance on visual cues means separation degrades if the source is fully occluded or leaves the camera's field of view. Future iterations could mitigate this by implementing audio-only fallback models when visual data is unavailable. Second, the computational cost of running concurrent detection, tracking, and separation models currently necessitates a tethered PC setup. Migrating MoXaRt to standalone headsets will require significant optimization, such as model quantization and on-device NPU acceleration. Another possible solution is to match the video see-through streaming to the processed audio with the same amount of constant latency; however, the effectiveness of such an approach regarding motion-to-photon latency requires further study.
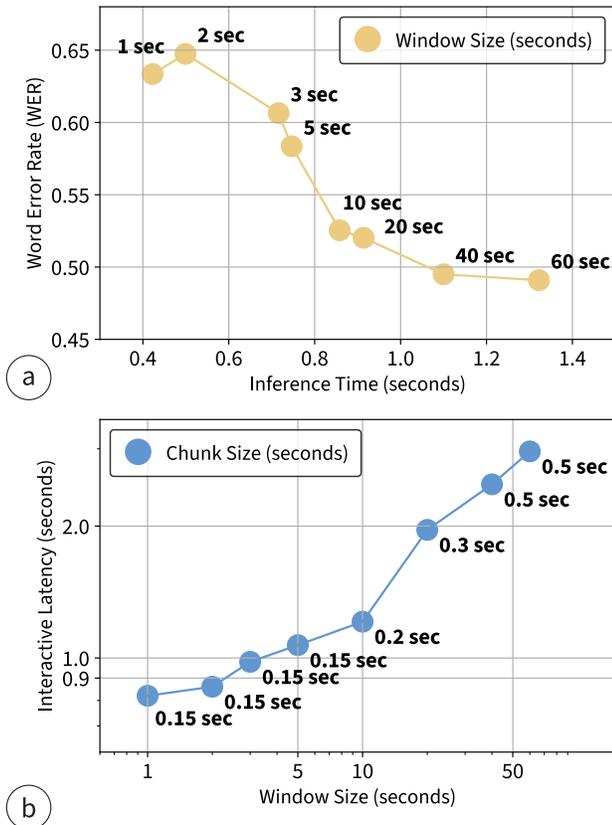
Figure 9: Model Performance vs. Real-time Analysis. (a) The size of the context window for model input (*window_size*) determines the trade-off between WER and inference time. (b) Reducing *chunk_size* can optimize interactive latency to as low as 0.82s, at the expense of audio separation quality.

7.3 Scalability and Persistence

A key challenge in deploying MoXaRt to the wild is handling the scale and dynamism of real-world acoustic environments. While our cascaded architecture supports dynamic source numbers, performance naturally degrades as the number of concurrent speakers increases beyond four [22]. We argue that the solution is not simply larger models, but reframing the problem as attention-driven separation [47]. Instead of attempting to separate all N sources simultaneously, future work should leverage gaze-based prioritization to isolate only the K sources relevant to the user, transforming an intractable signal processing task into a manageable HCI problem.

Furthermore, the system must maintain temporal consistency in these dynamic scenes. Our current window-based processing can be susceptible to "track swapping," where a speaker who is temporarily silent or occluded is assigned to a different audio channel upon reappearing. Future work will address this by integrating robust identification embeddings (*e.g., ArcFace*[13]) to anchor audio streams to specific individuals, ensuring that a user's volume adjustments persist throughout long-duration, multi-turn interactions.

7.4 Baseline and Interaction Design

In our human evaluation, we focused on comparing user experiences with and without MoXaRt. We omitted additional baselines because, to our knowledge, no existing system supports real-time, interactive control over a user's authentic soundscape. As discussed in Sec. 5, state-of-the-art separation models are typically designed for offline processing or restricted to specific domains, such as speech-only, lacking the throughput necessary for live interaction. Similarly, commercial noise-canceling headsets only filter audio by broad categories (*e.g.,* stationary noise versus speech) rather than isolating individual sound sources. These devices do not support the granular remixing enabled by our system and are incapable of the core interaction in our study: the selective amplification of a specific target within a multi-source environment. Consequently, they were excluded as comparable baselines.

However, we recognize that our current interaction design, which relies on handheld controllers, represents a limitation for social XR contexts. Future iterations of MoXaRt could benefit from more natural input modalities to provide a more seamless user experience. Potential avenues for future research include gaze-dwelling for source selection, micro-gestures for precise volume control, or neural interfaces (EMG/EEG) for discrete, hands-free interaction [46, 68].

7.5 Ethical and Social Implications

Beyond technical hurdles mentioned in the previous sections, the capacity to selectively filter the acoustic environment introduces complex social and ethical dynamics. Unlike noise-canceling headphones, which signal global isolation, AR-based selective filtering is invisible to bystanders. This raises questions regarding consent and shared reality: does a user have the ethical right to tune out a specific person in a collaborative setting, or silence a specific instrument in a live orchestra, without the source's knowledge or consent? Such capabilities fundamentally alter the social contract of presence. As these technologies mature, we must consider design frameworks that preserve social transparency, perhaps through external signaling (*e.g.,* a busy indicator on the headset) or protocols that balance user agency with bystander awareness.

8 Conclusion

In this paper, we presented MoXaRt, the first XR system to integrate real-time, audio-visual sound separation as a core primitive for direct, interactive control over a user's authentic soundscape. Our cascaded audio-visual transformer model leverages visual anchors (faces and instruments) to perform robust, real-time separation of individual sources from a single-channel input. Our comprehensive evaluation, performed on a new, complex audio-visual dataset we contributed for this task, confirmed our model's state-of-the-art technical performance. Furthermore, our 22-participant user study confirmed that MoXaRt significantly improves communication clarity, boosting listening comprehension by 36.2% ($p = 0.0058$), reduces cognitive load, and enhances the overall user experience with a new sense of interactive sound control.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. COCO-Stuff: Thing and Stuff Classes in Context. arXiv:1612.03716 [cs.CV] https://arxiv.org/abs/1612.03716

[2] Jean-François Cardoso. 1998. Blind Signal Separation: Statistical Principles. *Proc. IEEE* 86, 10 (1998), 2009–2025. doi:10.1109/5.720250

[3] Ruei-Che Chang, Chia-Sheng Hung, Bing-Yu Chen, Dhruv Jain, and Anhong Guo. 2024. SoundShift: Exploring Sound Manipulations for Accessible Mixed-Reality Awareness. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 116–132. doi:10.1145/3643834.3661556

[4] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M. Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement *(MobiSys '22)*. Association for Computing Machinery, New York, NY, USA, 384–396. doi:10.1145/3498361.3538933

[5] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. 2022. iQuery: Instruments as Queries for Audio-Visual Sound Separation. arXiv:2212.03814 [cs.CV] https://arxiv.org/abs/2212.03814

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. arXiv:1802.02611 [cs.CV] https://arxiv.org/abs/1802.02611

[7] Tuochao Chen, Qirui Wang, Runlin He, and Shyamnath Gollakota. 2025. Spatial Speech Translation: Translating Across Space With Binaural Hearables. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 352, 19 pages. doi:10.1145/3706598.3713745

[8] E. Colin Cherry. 1953. Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America* 25, 5 (1953), 975–979. doi:10.1121/1.1907229

[9] Hyunsung Cho, Naveen Sendhilnathan, Michael Nebeling, Tianyi Wang, Purnima Padmanabhan, Jonathan Browder, David Lindlbauer, Tanya R. Jonker, and Kashyap Todi. 2024. SonoHaptics: An Audio-Haptic Cursor for Gaze-Based Object Selection in XR. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 125, 19 pages. doi:10.1145/3654777.3676384

[10] Hyunsung Cho, Alexander Wang, Divya Kartik, Emily Liying Xie, Yukang Yan, and David Lindlbauer. 2024. Auptimize: Optimal Placement of Spatial Audio Cues for Extended Reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 126, 14 pages. doi:10.1145/3654777.3676424

[11] Jung-Woo Choi, Franz Zotter, Byeongho Jo, and Jae-Hyoun Yoo. 2022. Multiarray Eigenbeam-ESPRIT for 3D Sound Source Localization With Multiple Spherical Microphone Arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), 2310–2325. doi:10.1109/TASLP.2022.3183929

[12] Gustavo Corrêa De Almeida, Vinicius Costa de Souza, Luiz Gonzaga Da Silveira Júnior, and Maurício Roberto Veronez. 2024. Spatial Audio in Virtual Reality: A systematic review. In *Proceedings of the 25th Symposium on Virtual and Augmented Reality* (Rio Grande, Brazil) *(SVR '23)*. Association for Computing Machinery, New York, NY, USA, 264–268. doi:10.1145/3625008.3625042

[13] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. 2022. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (Oct. 2022), 5962–5979. doi:10.1109/tpami.2021.3087709

[14] derronqi. 2024. *yolov8-face: YOLOv8 face detection with landmarks.* https://github.com/derronqi/yolov8-face GitHub repository, commit 18f9fde.

[15] Henghui Du, Guangyao Li, Chang Zhou, Chunjie Zhang, Alan Zhao, and Di Hu. 2025. Crab: A Unified Audio-Visual Scene Understanding Model with Explicit Cooperation. arXiv:2503.13068 [cs.CV] https://arxiv.org/abs/2503.13068

[16] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics* 37, 4 (July 2018), 1–11. doi:10.1145/3197517.3201357

[17] Stefan Fichna, Thomas Biberger, Bernhard U. Seeber, and Stephan D. Ewert. 2021. Effect of Acoustic Scene Complexity and Visual Scene Representation on Auditory Perception in Virtual Audio-Visual Environments. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. 1–9. doi:10.1109/I3DA48870.2021.9610916

[18] Ruohan Gao and Kristen Grauman. 2021. VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency. In *CVPR*.

[19] Google. 2025. Gemini. https://gemini.google.com/.

[20] Gabriel Haas, Evgeny Stemasov, Michael Rietzler, and Enrico Rukzio. 2020. Interactive Auditory Mediated Reality: Towards User-defined Personal Soundscapes *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 16 pages. doi:10.1145/3357236.3395493

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531 [stat.ML] https://arxiv.org/abs/1503.02531

[22] Chao Huang, Susan Liang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. 2024. High-Quality Visually-Guided Sound Separation from Diverse Categories. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 35–49.

[23] Aapo Hyvärinen and Erkki Oja. 2000. Independent Component Analysis: Algorithms and Applications. *Neural Networks* 13, 4-5 (2000), 411–430. doi:10.1016/S0893-6080(00)00026-5

[24] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, and Andrew Fitzgibbon. 2011. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology* (Santa Barbara, California, USA) *(UIST '11)*. Association for Computing Machinery, New York, NY, USA, 559–568. doi:10.1145/2047196.2047270

[25] Dhruv Jain, Khoa Huynh Anh Nguyen, Steven Goodman, Rachel Grossman-Kahn, Hung Ngo, Aditya Kusupati, Ruofei Du, Alex Olwal, Leah Findlater, and Jon E. Froehlich. 2022. ProtoSound: A Personalized and Scalable Sound Recognition System for Deaf and Hard of Hearing Users. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16. doi:10.1145/3491102.3502020

[26] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. 2023. *Ultralytics YOLO.* https://github.com/ultralytics/ultralytics

[27] Ilya Kavalerov, Scott Wisdom, Hakan Erdogan, Brian Patton, Kevin Wilson, Jonathan Le Roux, and John R. Hershey. 2019. Universal Sound Separation. arXiv:1905.03330 [cs.SD] https://arxiv.org/abs/1905.03330

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV] https://arxiv.org/abs/2304.02643

[29] Wei-Tsung Lu, Ju-Chiang Wang, Qiuqiang Kong, and Yun-Ning Hung. 2024. Music Source Separation With Band-Split Rope Transformer. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 481–485. doi:10.1109/ICASSP48485.2024.10446843

[30] Joanna Luberadzka, Enric Gusó Muñoz, Umut Sayin, and Adan Garriga. 2025. Audio technology for improving social interaction in extended reality. *Frontiers in Virtual Reality* Volume 5 - 2024 (2025). doi:10.3389/frvir.2024.1442774

[31] Meta Reality Labs Research. 2022. Audio Presence: The "perceptual superpower" that will make the metaverse feel real. https://tech.facebook.com/reality-labs/2022/06/audio-presence-the-perceptual-superpower-that-will-make-the-metaverse-feel-real. Meta Tech Blog.

[32] Nikolaos Moustakas, Andreas Floros, Emmanouel Rovithis, and Konstantinos Vogklis. 2021. Prediction and Controlling of Auditory Perception in Augmented Environments. A Loudness-Based Dynamic Mixing Technique. *Applied Sciences* 11, 22 (2021). doi:10.3390/app112210944

[33] Rajeev Nongpiur and Tianyu Xu. 2025. Passively measuring room reverb. U.S. Patent Application No. 18/739,899. Published December 11, 2025.

[34] Rajeev Nongpiur and Tianyu Xu. 2025. System for measuring room reverb. U.S. Patent Application No. 18/627,621. Published October 9, 2025.

[35] Peter Ochieng. 2023. Deep neural network techniques for monaural speech enhancement: state of the art analysis. arXiv:2212.00369 [cs.SD] https://arxiv.org/abs/2212.00369

[36] OpenAI. 2025. ChatGPT. https://chat.openai.com/.

[37] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 741–754. doi:10.1145/2984511.2984517

[38] Katharine Patterson, Kevin Wilson, Scott Wisdom, and John R. Hershey. 2022. Distance-Based Sound Separation. arXiv:2207.00562 [cs.SD] https://arxiv.org/abs/2207.00562

[39] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. 2023. OpenScene: 3D Scene Understanding with Open Vocabularies. arXiv:2211.15654 [cs.CV] https://arxiv.org/abs/2211.15654

[40] Darius Petermann, Gordon Wichern, Zhong-Qiu Wang, and Jonathan Le Roux. 2022. The Cocktail Fork Problem: Three-Stem Audio Separation for Real-World Soundtracks. arXiv:2110.09958 [eess.AS] https://arxiv.org/abs/2110.09958

[41] Chandan K A Reddy, Vishak Gopal, and Ross Cutler. 2021. DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality metric to evaluate Noise Suppressors. arXiv:2010.15258 [cs.SD] https://arxiv.org/abs/2010.15258

[42] Simon Rouard, Francisco Massa, and Alexandre Défossez. 2022. Hybrid Transformers for Music Source Separation. arXiv:2211.08553 [eess.AS] https://arxiv.org/abs/2211.08553

[43] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. 2018. SDR - half-baked or well done? arXiv:1811.02508 [cs.SD] https://arxiv.org/abs/1811.02508

[44] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2019. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv:1801.04381 [cs.CV] https://arxiv.org/abs/1801.04381

[45] Laura Schütz, Sasan Matinfar, Ulrich Eck, Daniel Roth, and Nassir Navab. 2025. Sonify Anything: Towards Context-Aware Sonic Interactions in AR. In *Proceedings of the 2025 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. arXiv:2508.01789.

[46] Dongeek Shin and Tianyu Xu. 2025. Context-based user input control of near-eye displays. U.S. Patent Application No. 18/669,279. Published November 20, 2025.

[47] Dongeek Shin, Tianyu Xu, and Jennifer Irturk. 2025. Saliency-based mixed-order ambisonics encoding of multiple sound sources for immersive spatial audio. U.S. Patent No. 12,494,214. Issued December 9, 2025.

[48] Xia Su, Jon E. Froehlich, Eunyee Koh, and Chang Xiao. 2024. SonifyAR: Context-Aware Sound Generation in Augmented Reality. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 128, 13 pages. doi:10.1145/3654777.3676406

[49] John Sweller. 1988. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science* 12, 2 (1988), 257–285. doi:10.1207/s15516709cog1202_4

[50] Jiwen Tang. 2023. Research on the Application of Sound in Virtual Reality. *Highlights in Science, Engineering and Technology* 44 (04 2023), 206–212. doi:10.54097/hset.v44i.7323

[51] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: the new data in multimedia research. *Commun. ACM* 59, 2 (2016). doi:10.1145/2812802

[52] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel P. W. Ellis, and John R. Hershey. 2021. Into the Wild with AudioScope: Unsupervised Audio-Visual Separation of On-Screen Sounds. arXiv:2011.01143 [cs.SD] https://arxiv.org/abs/2011.01143

[53] Efthymios Tzinis, Scott Wisdom, Tal Remez, and John R. Hershey. 2022. AudioScopeV2: Audio-Visual Attention Architectures for Calibrated Open-Domain On-Screen Sound Separation. arXiv:2207.10141 [cs.SD] https://arxiv.org/abs/2207.10141

[54] Bandhav Veluri, Justin Chan, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota. 2023. Real-Time Target Sound Extraction. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49357.2023.10094573

[55] Bandhav Veluri, Malek Itani, Justin Chan, Takuya Yoshioka, and Shyamnath Gollakota. 2023. Semantic Hearing: Programming Acoustic Scenes with Binaural Hearables. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. 1–14. doi:10.1145/3586183.3606779

[56] Bandhav Veluri, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota. 2024. Look Once to Hear: Target Speech Hearing with Noisy Examples. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 37, 16 pages. doi:10.1145/3613904.3642057

[57] Helin Wang, Jiarui Hai, Yen-Ju Lu, Karan Thakkar, Mounya Elhilali, and Najim Dehak. 2025. SoloAudio: Target Sound Extraction with Language-oriented Audio Diffusion Transformer. arXiv:2409.08425 [eess.AS] https://arxiv.org/abs/2409.08425

[58] Tianyi Wang, Xun Qian, Fengming He, Xiyun Hu, Yuanzhi Cao, and Karthik Ramani. 2021. GesturAR: An Authoring System for Creating Freehand Interactive Augmented Reality Applications. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 552–567. doi:10.1145/3472749.3474769

[59] Scott Wisdom, Efthymios Tzinis, Hakan Erdogan, Ron J. Weiss, Kevin Wilson, and John R. Hershey. 2020. Unsupervised Sound Separation Using Mixture Invariant Training. arXiv:2006.12701 [eess.AS] https://arxiv.org/abs/2006.12701

[60] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. doi:10.1145/1978942.1978963

[61] Brandon Woodard, Margarita Geleta, Joseph J. LaViola Jr., Andrea Fanelli, and Rhonda Wilson. 2025. AudioMiXR: Spatial Audio Object Manipulation with 6DoF for Sound Design in Augmented Reality. arXiv:2502.02929 [cs.HC] https://arxiv.org/abs/2502.02929

[62] Muzhe Wu, Yi Fei Cheng, and David Lindlbauer. 2024. New Ears: An Exploratory Study of Audio Interaction Techniques for Performing Search in a Virtual Reality Environment. In *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 386–395. doi:10.1109/ISMAR62088.2024.00053

[63] Tianyu Xu, Jihan Li, Penghe Zu, Pranav Sahay, Maruchi Kim, Jack Obeng-Marnu, Farley Miller, Xun Qian, Katrina Passarella, Mahitha Rachumalla, Rajeev Nongpiur, and D Shin. 2025. Enhancing XR Auditory Realism via Multimodal Scene-Aware Acoustic Rendering *(UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 17, 16 pages. doi:10.1145/3746059.3747730

[64] Jing Yang, Amit Barde, and M. Billinghurst. 2022. Audio Augmented Reality: A Systematic Review of Technologies, Applications, and Future Research Directions. *Journal of the Audio Engineering Society* (2022). doi:10.17743/jaes.2022.0048

[65] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. 2018. The Sound of Pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 570–586. doi:10.1007/978-3-030-01246-5_35

[66] Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jiaqi Yip, Dianwen Ng, and Bin Ma. 2024. MossFormer2: Combining Transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation. arXiv:2312.11825 [cs.SD] https://arxiv.org/abs/2312.11825

[67] Shengkui Zhao, Zexu Pan, and Bin Ma. 2025. ClearerVoice-Studio: Bridging Advanced Speech Processing Research and Practical Deployment. arXiv:2506.19398 [cs.SD] https://arxiv.org/abs/2506.19398

[68] Junyi Zhu, Tianyu Xu, Jiayu Wang, Emily Guan, JaeYoung Moon, Stiven Morvan, D Shin, Andrea Colaço, Stefanie Mueller, Karan Ahuja, Yiyue Luo, and Ishan Chatterjee. 2025. EI-Lite: Electrical Impedance Sensing for Micro-gesture Recognition and Pinch Force Estimation. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25)*. Association for Computing Machinery, New York, NY, USA, Article 23, 14 pages. doi:10.1145/3746059.3747671

A Technical Evaluation Dataset Collection Setup

A.1 Recording Environments

We recorded our datasets in the following two acoustic environments:

(1) A small study room (approx. 120 sqft) was used for scenarios C1–C5, C10, and C6-1.
(2) A large music school classroom containing a piano (approx. 1500 sqft) was used for scenarios C7–C9, C6-2, and C6-3.

A.2 Recording Hardware

Figure 10: Data collection setup: A Meta Quest 3 with 2 attached Rode Wireless GO II microphones captures the mono audio, while a stationary iPhone records a third-person view of the scene.

We employed a dual-system setup for data capture, as depicted in Figure 10. The primary system was designed to capture a stable, user-centric perspective using a Meta Quest 3 headset mounted on a stand at approximately head-level. To record the primary mono audio that serves as the input to our system, a Rode Wireless GO II dual-channel microphone system was attached to the headset, and set to mono-mode to mimic the auditory input via microphones on head mounted devices. Simultaneously, a second system—a stationary iPhone (iPhone 13 Pro, or an iPhone 15 Pro for scenarios C8 and C9) on a tripod—recorded a third-person view of the entire scene for video reference.

A.3 Speech Content Generation

The spoken content for our recordings was sourced to reflect realistic use cases. Single-person speech segments were transcribed from CNN Student News broadcasts. Two-speaker dialogues were generated using ChatGPT (gpt-4o) with prompts such as, "Could you help me generate some simulated TOEFL-style conversations between two people, each about one minute long?" The generated content spanned daily conversations, academic discussions, and professor-student interactions.

A.4 Data Synchronization

To ensure precise alignment between the centrally recorded video stream and the multiple independent ground truth audio streams, we employed a manual synchronization method. At the start of each take, a single, sharp clap was performed in the scene. In post-processing, we used Adobe Premiere Pro to align all audio and video tracks by matching the sharp transient peak of the clap event across all recordings. This technique is analogous to the function of a traditional clapperboard in film production.

B DAVIS Baseline Training Details

This appendix details the training procedure, evaluation preprocessing, and qualitative performance analysis of the DAVIS baseline model used in our experiments. As our evaluation scenarios involve three concurrent sound sources, we retrained the original DAVIS model, which was designed for two-source separation, to support three-source separation.

B.1 Model Training

We trained a 3-source separation model based on the DAVIS framework, specifically using the improved flow-matching variant, DAVIS_Flow, which demonstrated superior performance in our preliminary tests.

- **Architecture Choice:** We trained both the original DAVIS model and the DAVIS_Flow variant for 3-source separation on the MUSIC dataset. The DAVIS_Flow model converged to a validation Signal-to-Distortion Ratio (SDR) of approximately 5, whereas the original DAVIS model only reached an SDR of 2. Consequently, we selected the DAVIS_Flow model for all subsequent experiments and evaluations.
- **Dataset:** We used the MUSIC dataset provided by the official DAVIS repository. The training split remained unchanged. For validation, we generated new 3-source mixtures by first extracting a complete list of non-overlapping audio files from the provided metadata and then randomly sampling and mixing three files from distinct instrument classes, mirroring the logic of the training data loader.
- **Hardware:** The model was trained on a single NVIDIA A40 GPU (40GB VRAM, CUDA 12.8) with 128GB of main memory and 8 CPU cores.
- **Hyperparameters:** We adopted the hyperparameters from the official DAVIS repository with the following necessary modifications for our 3-source, single-GPU setup:
  - `-num_mix 3` (changed from 2)
  - `-num_gpus 1` and `-gpu_ids 0` (changed from a multi-GPU setup)
  - `-batch_size_per_gpu 6` (increased from 4 to utilize available memory)
- **Training Time:**
  - **On MUSIC dataset:** Training the 3-source DAVIS_Flow model from scratch took approximately four days to converge, at a rate of 3–4 epochs per hour.
  - **On AVE dataset:** Due to the significantly longer training time on AVE (approx. 1.5 hours per epoch), we fine-tuned the publicly available 2-source pre-trained checkpoint for 3-source separation. The training took about three days, though validation metrics fluctuated without showing significant improvement over the course of training.

B.2 Evaluation Preprocessing

To prepare our evaluation scenarios for the DAVIS_Flow baseline, we followed a multi-step preprocessing pipeline.

(1) **Segmentation:** We first segmented each one-minute video into smaller, overlapping chunks using the provided shell scripts. For scenarios involving the MUSIC dataset, we used an audio length (audLen) of 66,150 at a sample rate (audRate) of 11,025 Hz, resulting in approximately 6 segments per video. For scenarios involving the AVE dataset, we used an audLen of 110,000 at an audRate of 11,000 Hz, resulting in about 10 segments per video.
(2) **Model Inference:** We ran the appropriate pre-trained DAVIS_Flow model on the segmented data corresponding to each evaluation scenario.
(3) **Concatenation:** The separated audio segments from the model were concatenated to reconstruct the full-length, one-minute audio tracks.
(4) **Manual Alignment:** Due to the segmentation and concatenation process, minor length mismatches occurred in some outputs. We used Adobe Premiere Pro to manually trim repeated audio artifacts in four of the output files (C5-1, C8-3, C9-1, C9-3) to ensure their length precisely matched the one-minute input.

B.3 Qualitative Performance Analysis

Our evaluation revealed clear patterns in the model's performance, highlighting its strengths and weaknesses.

- **Cases of Poor Performance:**
  - **Speaker Separation:** The model generative perform poorly on speaker separation tasks. The output sometimes degrade into white noise, which explains its worse-than-baseline performance on WER, as shown in Table 2. We hypothesize that this is because the AVE dataset, on which the speech separation model was trained, uses broad class labels like "Male speech" or "Female speech" and is not designed for separating multiple instances of the same class (*e.g.*, two different male speakers).
  - **Piano Separation:** The model also performed poorly when a piano was one of the sound sources. The separated track for the piano was either a distorted copy of another instrument's audio or white noise. This is an expected failure, as the MUSIC dataset used for training does not contain a 'piano' instrument class, making this an out-of-distribution task.
- **Cases of Good Performance:**
  - The model performed well on two-source musical instrument separation tasks, provided the piano was not one of the sources. It also successfully separated flute and guitar sounds even when accompanied by human speech, demonstrating a degree of robustness in handling mixtures of music and speech.