ASAudio: A Survey of Advanced Spatial Audio Research

Zhiyuan Zhu* **Yu Zhang*** **Wenxiang Guo*** **Changhao Pan*** **Zhou Zhao†**
Zhejiang University
{schmittzhu, zhaozhou}@zju.edu.cn

Abstract

With the rapid development of spatial audio technologies today, applications in AR, VR and other scenarios have garnered extensive attention. Unlike traditional mono sound, spatial audio offers a more realistic and immersive auditory experience. Despite notable progress in the field, there remains a lack of comprehensive surveys that systematically organize and analyze these methods and their underlying technologies. In this paper, we provide a comprehensive overview of spatial audio and systematically review recent literature in the area. To address this, we chronologically outline existing work related to spatial audio and categorize these studies based on input-output representations, as well as generation and understanding tasks, thereby summarizing various research aspects of spatial audio. In addition, we review related datasets, evaluation metrics, and benchmarks, offering insights from both training and evaluation perspectives. Related materials are available at https://github.com/dieKarotte/ASAudio.

1 Introduction

Spatial audio delivers an immersive, three-dimensional listening experience by simulating how sound propagates and is perceived in space, representing the culmination of audio's evolution from mono to surround sound (Poeschl et al., 2013). Fueled by its adoption as a core feature in products from Apple, Google, and Meta, the technology has seen accelerated development and widespread application in film, gaming, and the emerging metaverse (Chen et al., 2025; Wuolio and Moreira Kares, 2023; Lee et al., 2023a; Broderick et al., 2018; Murphy and Neff, 2011), which in turn has sharpened the focus of academic research.

As illustrated in Fig. 1, the research landscape of spatial audio has undergone a significant evolution. Before 2021, efforts primarily center on understanding tasks like sound event localization and detection (SELD) and source separation, dominated by foundational CNN-based models (Zhou et al., 2020; Gao and Grauman, 2019; Wu et al., 2021; Richard et al., 2021; Nguyen et al., 2022; Shimada et al., 2021) and limited by the scale of early datasets (Donley et al., 2021; Morgado et al., 2020). Since 2022, the field has entered a new phase of rapid, synergistic advancement in both understanding and generation, fueled by breakthroughs in generative models and the proliferation of multimodal datasets (Zheng et al., 2024; Zhang et al., 2025; Kim et al., 2025; Sun et al., 2024). This period sees the rise of powerful generation models like ImmerseDiffusion (Heydari et al., 2025) and Diff-SAGe (Kushwaha et al., 2025), which drastically improves audio quality and realism. Crucially, the underlying technologies, such as attention mechanisms and large language models, also revolutionize understanding. This propels the task from traditional signal-level analysis toward higher-level semantic reasoning, as seen in advanced models for attention-based separation (Ye et al., 2024) and LLM-based spatial inference (Zheng et al., 2024).

To systematically review these advances in representation, understanding, generation, datasets, and evaluation protocols, this paper is organized as follows: Section 2 discusses input-output representations, Sections 3 and 4 analyze understanding and generation tasks, and Section 5 summarizes existing datasets and evaluation standards.

2 Representations of Spatial Audio

2.1 Inputs Representations

Input representations aim to capture semantic, acoustic, and spatial information. They are provided alone or in combination as mono audio, text, visual signals, or spatial coordinates. We provide a detailed explanation of input representation and

*Equal contribution
†Corresponding Author

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 417–442

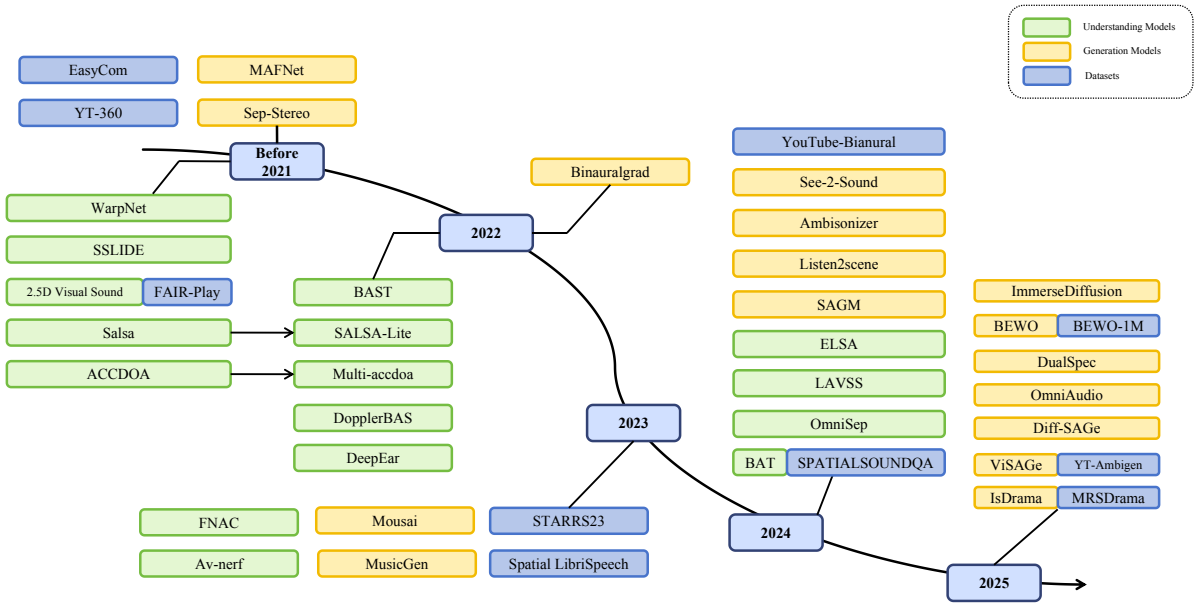

Figure 1: A timeline of spatial audio models & datasets in recent years. The timeline is established mainly according to the release date of the technical paper for each model. We mark the understanding models in green and the generation models in yellow, while datasets are marked in blue. Arrows indicate the evolution of models.

their primary processing method in Figure 2.

Natural Language Prompts Natural language prompts specify semantic content and spatial attributes in an intuitive way. They describe events for generation (Kreuk et al., 2022; Liu et al., 2023) or serve as queries in understanding tasks. For example, BAT (Zheng et al., 2024) uses a large language model to process question–answer pairs about sound event detection, direction estimation, and spatial reasoning, and it extracts spatial information from natural language.

Spatial Position Explicit spatial position data, such as Cartesian or spherical coordinates, provides direct guidance to place sources in generation tasks and serves as ground truth for localization models in understanding tasks. Some studies (Liu et al., 2022; Zhang et al., 2025) also include radial velocity and orientation. They simulate Doppler effects to enhance dynamic properties.

Visual Information Visual information (images or videos) strongly correlates with sound and provides valuable spatial and semantic context. It offers key cues for audio–visual source separation and localization (Zhao et al., 2018; Ye et al., 2024; Zhou et al., 2018) and for audio–visual acoustic matching (Chen et al., 2022). It also guides mono-to-spatial generation (Gan et al., 2019; Gao and Grauman, 2019) and video-to-spatial-audio generation (Liu et al., 2025a) tasks.

Monoaural Audio Mono audio serves as the base acoustic content in many generation tasks. It supplies core timbral and spectral cues. In two-stage systems, the mono stream is first processed and then "upmixed" into multichannel or binaural formats under the guidance of spatial inputs such as visuals or positions information.

2.2 Spatial Cues and Physical Modeling

A core aspect of spatial audio is the accurate modeling of sound propagation and perception in three-dimensional space. On the other hand, hardware and human hearing background are important parts in spatial audio. We will introduce two key concepts including the room impulse response (RIR) and the head-related transfer function (HRTF). Also, recording hardware like multi-channel microphones and how humans perceive sound localization are introduced.

Room Impulse Response (RIR) The room impulse response (RIR) characterizes all acoustic paths from a source to a receiver, bridging virtual and real acoustics. As direct measurement is costly, research has focused on alternatives. Some

methods estimate RIRs from visual inputs to avoid acoustic measurements (Kim et al., 2019; Ratnarajah et al., 2024; Majumder et al., 2022), while others use simulation tools to generate data for training and improving tasks like source separation (Roman et al., 2024; Ahn et al., 2023; Jeub et al., 2009; Vacher et al., 2014; Mittag et al., 2017; Di Carlo et al., 2021; Grondin et al., 2020; Xu et al., 2021). To support complex applications, precise RIRs have been measured for specific scenarios like dense grids or dynamic sources (Koyama et al., 2021; Ratnarajah et al., 2022; Politis et al., 2020; McKenzie et al., 2021b,a), and perceptual evaluation often relies on measured binaural RIRs (BRIRs) to assess synthesis authenticity (Brinkmann et al., 2017).

Binaural hearing and spatial cues Humans perceive three dimensional sound based on the binaural hearing system. The brain compares and analyzes the signals at the two ears to estimate source direction and distance. This process relies on a set of acoustic cues. Among them, interaural time difference (ITD) and interaural level difference (ILD) (Moore, 2012) are the two core physical quantities for horizontal localization. The classic Duplex Theory states that ITD is the main cue at low frequencies(roughly below 1.5 kHz) while ILD is more important at higher frequencies.

ITD arises from the path length difference of the wavefronts arriving at the two ears. When the source is off the median plane, the wave reaches the near ear first and then diffracts around the head to the far ear. The small delay, up to about 0.6–0.8 ms, is detected by the auditory system and encodes azimuth. ILD is caused by the head's acoustic shadow. For high frequencies with shorter wavelengths, the head blocks sound so that the far ear receives a weaker signal. The level difference becomes a key cue for high frequency localization.

Spatial audio hardware Spatial audio needs a capture-to-render hardware stack. Microphone arrays (Blanco Galindo et al., 2020) sample the 3D sound field by placing omnidirectional mics in designed geometries; the signals carry spatial cues. By source distance, arrays use near-field or far-field models. By topology they are linear, planar, or volumetric (Benesty et al., 2008).

As a special array, a dummy head microphone (Lübeck et al., 2022) places microphones at the ear canal entrances of a realistic head model. It directly reproduces human binaural hearing and natively records key cues such as ITD and ILD which provides highly realistic immersion for headphone playback.

Head-Related Transfer Function Head-related transfer function (HRTF) is a subject-specific filter describing how an individual's anatomy alters incoming sound, encoding the binaural and monaural cues essential for 3D perception. Because HRTFs are highly individualized, personalization is critical to avoid perceptual artifacts like in-head localization and front-back confusion. To this end, researchers have developed several methods. Some predict HRTFs from anthropometric features like ear shape using neural networks (Warnecke et al., 2022; Arbel et al., 2024; Zhao et al., 2022). Others select the best-matching HRTF from a database, guided by perception-aligned metrics (Lee et al., 2023b; Marggraf-Turley et al., 2024). The most mainstream approach, however, is spatial upsampling from sparse data, which uses deep models to interpolate a full HRTF from a few measurements. This includes using various deep architectures like CNNs and Transformers for reconstruction (Jiang et al., 2023; Ito et al., 2022; Hogg et al., 2024; Ma et al., 2023; Zhang et al., 2023), incorporating physical priors to improve performance (Chen et al., 2023; Thuillier et al., 2024), and leveraging neural fields to represent HRTFs as continuous functions (Zhang et al., 2023; Masuyama et al., 2024). Future work aims to fuse these methods and deploy them on consumer devices (Warnecke et al., 2022; Jiang et al., 2023).

2.3 Output Representations

Spatial audio is mainly represented in three formats. Channel-based formats (e.g., 5.1 or 7.1 surround) assign signals to predefined loudspeaker positions. Scene-based formats (e.g., higher-order Ambisonics (HOA)) represent the full three-dimensional sound field using spherical harmonic decomposition. Object-based formats, such as Dolby Atmos, treat each source as an independent object with positional metadata and render it dynamically at playback. We analyze three output paradigms and discuss binaural rendering separately.

Channel-Based Audio Channel-based audio maps signals to predefined loudspeaker positions, such as stereo, 5.1, or 7.1. Spatial position is implied by level and time differences across channels. The psychoacoustic basis is summing localization.

Amplitude panning follows the sine law:

$$\sin\theta_I = \frac{E_L - E_R}{E_L + E_R}\sin\theta_0. \quad (1)$$

This paradigm is widely used but depends on standardized layouts. It has a small "sweet spot" and limited flexibility and scalability.

Scene-Based Audio Scene-based audio aims to capture and physically reproduce the entire sound field within a region. Key methods include Ambisonics (Zotter and Frank, 2019) and wave field synthesis (WFS). Ambisonics represents the 3D field by spherical harmonic decomposition:

$$P(\mathbf{x},\omega) = \sum_{n=0}^{N}\sum_{m=-n}^{n} \alpha_n^m(\omega)\, j_n(kr)\, Y_n^m(\hat{\mathbf{x}}). \quad (2)$$

This paradigm produces a wide and stable listening area. However, it places high demands on the system, which limits adoption in the consumer market. It decomposes the field into spherical harmonics (Malham and Myatt, 1995) for a device independent description and offers a wide, stable listening area as the listener moves. First order ambisonics (FOA) uses four B format channels. W is an omnidirectional component that represents overall sound pressure and ambient impression. X, Y, and Z are figure of eight components aligned with the Cartesian axes and encode front back, left right, and up down sound energy. Higher order ambisonics (HOA) provide finer spatial resolution.

Object-Based Audio Object-based audio treats each source as an independent audio object that carries content and metadata, such as position and trajectory. The final mix is rendered in real time on the playback device. Dolby Atmos is a representative system. By decoupling content from the physical playback setup, this paradigm achieves strong scalability and interactivity and becomes a core of next-generation immersive media.

Binaural Audio Binaural audio is a key rendering method and the final form that delivers advanced spatial formats to the ears over headphones. It uses HRTFs to reconstruct the ear-canal pressure and thus tricks the brain into perceiving a 3D scene. Convincing experiences require dynamic head tracking and room acoustics (reverberation) modeling. These components reduce front–back confusion and promote externalization.

2.4 Representation Discussion

Input representations We observe three axes that govern design choices: (i) **Abstraction vs Control precision**. Natural language and vision information are human-friendly and scalable for high-level intent, but suffer from ambiguity and lower precision; spatial coordinates deliver exact, reproducible control but lack semantics and are tedious to author. (ii) **Semantics vs Geometry**. High-level intents require an interpretation layer (often an LLM or structured parsers) to map semantics to machine-executable spatial parameters; geometric inputs bypass this layer but reduce expressivity. (iii) **Content vs Spatialization**. Monaural audio supplies core acoustic content (timbre, pitch), while other modalities guide spatial rendering; a two-stage pipeline (content generation, then spatialization) yields modularity and controllability. We make a concise comparison and discussion in Appendix A.1, Table 1.

Output representations The three output forms differ in device dependence, scalability, listening freedom, and playback-side complexity. Channel-based formats have high device dependence but low playback complexity. Scene-based formats offer high listening freedom but place strict demands on the system. Object-based formats provide unmatched flexibility and scalability and act as a core driver of next-generation immersive media. These paradigms are not mutually exclusive, and each suits different applications best. A concise comparison is deferred to Appendix A.2, Table 3. Most works adopt a single format output so a direct quantitative comparisons across output formats are rare. We will highlight this as an important direction in future work.

3 Understanding Approaches

Spatial audio understanding aims to analyze complex acoustic scenes by exploiting spatial cues. Core tasks include sound event localization and detection (SELD), spatial audio separation, and joint learning with visual and language modalities.

3.1 SELD Tasks

Sound event localization and detection (SELD) answers two questions at once: what sound occurs (sound event detection, SED) and where it comes from (direction of arrival estimation, DOAE). Traditional methods rely on signal processing while

modern work increasingly adopts deep learning models on SELD tasks.

Deep learning achieves strong progress on SELD with diverse network architectures. Early work (May et al., 2010) models binaural cues (ITD/ILD) with Gaussian mixtures to estimate azimuth and lays the foundation for later studies. SELDnet (Adavanne et al., 2018a) uses a CRNN to process SED and DOA in parallel and becomes a key baseline. To further improve performance, researchers explore alternative representations and mappings. For example, (Pavlidi et al., 2015) estimates the active intensity vector, while (Rana et al., 2019) builds an automated pipeline for Ambisonics estimation from audio–visual features. For binaural devices such as hearing aids, DeepEar (Yang and Zheng, 2022) designs a multi-sector network that localizes multiple sources. To handle unknown numbers of sources in the wild, (Kim et al., 2023) proposes a YOLO-inspired, event-driven localizer that is robust to concurrent events.

Jointly learning SED and DOA often degrades performance. Several strategies address this issue. (Cao et al., 2019) shows that two-stage training allows SED features to benefit DOAE. (Cao et al., 2021) introduces a track-wise output format, permutation-invariant training (PIT), and soft parameter sharing to avoid sacrificing subtask accuracy. (Shimada et al., 2021, 2022) proposes ACCDOA and its multi-target extension, which unify SELD as a single-target regression problem and remove the need to balance multi-task losses. SALSA (Nguyen et al., 2022) designs a joint time–frequency feature that maps signal energy and directional cues with high precision.

To fuse complementary strengths, (Yasuda et al., 2020) combines physics-based intensity vector (IV) estimation with DNN denoising and source separation to handle overlaps. With listener motion, (Krause et al., 2023) confirms the benefit of motion cues for localization, and (García-Barrios et al., 2022) analyzes how head rotations affect accuracy. In model design, self-supervised methods (Sun et al., 2023; Santos et al., 2024) and audio–visual learning (Gan et al., 2019; Tian et al., 2018) reduce dependence on large labeled sets. Recent architectures including CRNNs with SE modules (Naranjo-Alcazar et al., 2020), Transformers (Kuang et al., 2022), autoencoders (Huang et al., 2020; Wu et al., 2021), and VAEs (Bianco et al., 2020, 2021) capture time–frequency structure and support unsupervised or semi-supervised settings.

Recent works use diverse datasets and protocols, which complicates uniform comparison. Appendix B lists the datasets and metrics used in SELD tasks and report comparable performance.

3.2 Spatial Audio Separation

Source separation aims to recover individual sources from a mixture. With binaural or multichannel inputs, inter-channel spatial cues provide strong leverage, especially for challenging "cocktail party" scenarios.

Binaural Audio Separation Binaural separation uses ITD and ILD cues between the two ears to disentangle overlapping sources. Early machine learning approaches, such as (Weiss et al., 2009), employ probabilistic models. To support human–robot interaction, (Deleforge and Horaud, 2012) proposes a generative model with active binaural hearing so that a robot performs robust separation and localization in cocktail-party conditions. To handle multi-speaker separation under reverberation, (Zhang and Wang, 2017) introduces a novel 2D ITD feature, while (Wang and Wang, 2018) tightly integrates spectral and spatial features in a deep framework. To preserve spatial cues that matter to downstream applications, (Han et al., 2020) proposes MIMO TasNet for real-time speech separation with binaural cue retention.

Audio–visual fusion is another major direction. The pioneering 2.5D Visual Sound (Gao and Grauman, 2019) adopts a mix-and-separate strategy, where visual cues guide binaural separation. To go beyond systems that only model acoustics and ignore spatial position, LAVSS (Ye et al., 2024) introduces audio–visual spatial source separation (AVSS). It encodes object locations explicitly to steer the separation process.

Multichannel Audio Separation Multichannel separation uses richer spatial information and array geometry to address the underdetermined case where sources outnumber channels. Traditional methods such as spatial clustering (Wang et al., 2018) cluster time–frequency bins with GMMs using inter-channel cues (ITD, ILD, etc.). Early DNN work (Nugraha et al., 2016) combines DNN-modeled spectra with a classical multichannel Gaussian model to exploit spatial structure. Recent unsupervised methods, such as (Zmolikova et al., 2021), adopt variational Bayes to unify spectral and spatial cues and achieve end-to-end spatial separation. In addition, (Wang et al., 2018) proposes an

efficient algorithm that extends two-channel deep clustering to arbitrary microphone arrays. Similarly, (Morgado et al., 2018) converts mono audio to multichannel spatial audio via video analysis and implicitly separates and localizes unknown sources.

3.3 Cross-Modal Scene Understanding

To reach comprehensive scene understanding, spatial audio is increasingly learned together with other modalities, such as vision and natural language. The goal is to align and exploit the rich cues present across modalities.

Alignment Between Audio & Visual Information Aligning spatial audio with vision is key to cross-modal reasoning. (Morgado et al., 2020) propose audio–visual spatial alignment (AVSA) and use contrastive learning to capture correspondences between 360° videos and their spatial audio. (Yang et al., 2020) design a self-supervised task that asks the model to detect whether left–right audio channels are swapped. This task forces the model to learn spatial correspondence between audio modality and video modality.

Environment Information Understanding room acoustics is essential for realistic reproduction. (Liang et al., 2023) integrates propagation priors into NeRF to synthesize spatial audio consistent with novel views. (Luo et al., 2022) proposes neural acoustic fields (NAFs) that learn an implicit representation of sound propagation directly from impulse responses. Many studies (Savioja and Svensson, 2015; Ratnarajah et al., 2024; Bryan, 2020; ISO, 2009; Coldenhoff et al., 2024; Majumder et al., 2022; Srivastava et al., 2021) simulate or measure room impulse responses to analyze indoor acoustic parameters and capture geometry and material properties.

Visual Segmentation & Depth Estimation Depth and segmentation provide precise geometric supervision for spatial audio processing. (Liu et al., 2025b) integrates YOLOv8(Varghese and Sambath, 2024) detection with Depth Anything to estimate depth. It then computes accurate 3D source positions and supplies key cues for downstream spatialization.

Natural Language Guided Natural language guidance is a new frontier for spatial audio understanding. Because existing audio foundation models usually lack spatial awareness, ELSA (Devnani et al., 2024) uses contrastive learning and spatial regression targets to align spatial audio with text for the first time. BAT (Zheng et al., 2024) builds a new dataset, SPATIALSOUNDQA, with spatial question–answer pairs and fine-tunes a large language model (LLaMA-2). It shows the strong potential of LLMs for spatial audio reasoning.

3.4 Future Work

Spatial audio understanding may move from perception to cognition by incorporating explicit causal reasoning. Models may infer why events occur and what is likely to follow, rather than only identifying what and where. Spatial cues can serve as evidence for learning event chains, for example a glass falling that results in shattering. This requires unified multimodal foundation models that treat spatial audio as a first-class modality alongside vision and language which perform end-to-end reasoning instead of late fusion, enabling seamless cross-modal inference. The goal is joint outputs that capture 3D layout, event logic, and human dynamics in a single coherent representation. Progress also depends on learning counterfactuals and intervention effects grounded in basic physics. Benchmarking must evolve to test causal competence and embodied understanding, not just associative pattern matching.

4 Spatial Audio Generation Methods

Spatial audio generation evolves from traditional digital signal processing to advanced deep learning methods. This progress is driven by rapid advances in generative models. This section reviews recent developments, covering both cascade models and end-to-end models. A summary of recent deep learning models is presented in the Appendix C and Table 5 with their input/output format and model framework.

4.1 Cascade Models

This part focuses on a core topic in spatial audio generation: upmixing monaural audio into binaural audio with three-dimensional spatial cues. The "mono-to-binaural" process builds immersive listening. It aims to reproduce the spatial cues that humans perceive and traces the technical path from structured physical models to deep, especially vision-guided, frameworks.

Traditional Methods Humans localize sound with binaural hearing. This mechanism involves an ITD, an ILD, and spectral changes described by

HRTF. Early work such as (Brown and Duda, 1998) explicitly models wave propagation and diffraction with a simplified time-domain description. The model is interpretable and efficient. With deep learning, (Richard et al., 2021) introduces a neural rendering network that synthesizes binaural waveforms from a mono input and the listener position. The work shows the limits of a plain L_2 loss on raw waveforms.

Visually Guided Audio Spatialization A mono signal lacks spatial location information. Visually guided spatialization uses synchronized video to provide key context. The pioneering 2.5D Visual Sound framework (Gao and Grauman, 2019) employs a deep convolutional network to recover spatial cues and sets the basic paradigm. (Li et al., 2024c) adds object-level visual cues and designs a cyclic locate-and-upmix (CLUP) framework. It jointly learns visual source localization and binaural generation. To improve accuracy, researchers add 3D geometry. (Parida et al., 2022) stresses depth maps and designs an encoder–decoder with hierarchical attention. (Garg et al., 2021) separates geometry cues with a multi-task network and learns geometry-aware features. Efficient cross-modal fusion becomes a focus. (Zhang and Shao, 2021) proposes the multi-attention fusion network (MAFNet). (Liu et al., 2024) adds a novel audio–visual matching loss. (Zheng et al., 2022) defines a "binaural ratio" linked to physical cues to improve interpretability. (Li et al., 2024b) introduces a GAN framework with shared visual guidance and proposes a new spatial metric.

Audio Quality Enhancement After solving localization, another line improves audio fidelity and physical realism. (Leng et al., 2022) first applies diffusion. It generates shared and ear-specific information in two stages. (Liu et al., 2022) adds a plug-and-play DopplerBAS module that uses radial velocity to handle Doppler effects. (Lee and Lee, 2023) proposes the Neural Fourier Shift (NFS) network, which renders in the Fourier domain and predicts early reflections, cutting computation.

Weakly-Supervised/Self-Supervised Paradigms To break data limits, researchers propose new learning paradigms. (Xu et al., 2021) creates PseudoBinaural. It uses physical priors to make pseudo labels from many mono videos. (Rachavarapu et al., 2021) uses source localization as a proxy task for weak supervision. Multi-task and self-supervised learning also help. Sep-Stereo (Zhou et al., 2020) adds visual-guided separation as a second task. (Lin and Wang, 2021) enforces left–right consistency. (Li et al., 2021) adds a channel-flip classification task for self-supervision.

4.2 End-to-End Models

End-to-end spatial audio generation no longer upmixes an existing mono track. It synthesizes a complete sound field from high-dimensional, multimodal inputs such as silent video, natural language, or 3D geometry. The rise of diffusion models, large multimodal datasets, and cross-modal representation learning (e.g., CLIP) drives this paradigm. Early systems include the VQ-VAE framework in (Huang et al., 2022) and the surround-to-binaural network in (Yang et al., 2022).

Video-Driven Spatial Audio Generation The video-driven generation paradigm turns AI from a post-production tool into a creative engine. ViSAGe (Kim et al., 2025) generates first-order Ambisonics (FOA) from silent video and surpasses cascade methods. With VR/AR, generating immersive audio for 360° videos becomes important. OmniAudio (Liu et al., 2025a) tackles the 360V2SA task with a dual-branch design that uses panoramic and normal views. Other work (Rana et al., 2019; Liang et al., 2023) estimates 3D source positions from audio–visual cues and encodes them in panoramic sound.

Text and Multimodal Conditioned Generation Controlling spatial audio with natural language is a cutting-edge direction. Diffusion models drive this change. (Heydari et al., 2025) uses a latent diffusion model to produce 3D immersive soundscapes from text. It supports descriptive and parametric control. (Sun et al., 2024) notes that plain text embeddings blur spatial cues. It proposes SpatialSonic, which adds a spatial encoder and an azimuth-elevation matrix for explicit guidance. Architectural innovation then improves controllability. DualSpec (Zhao et al., 2025) introduces a pretrained separator and a channel-shift loss to enhance spatialization. Other studies, such as (Kushwaha et al., 2025; Zang et al., 2024), generate FOA from class labels and positions or directly from text. The trend extends to complex dialog and music. ISDrama (Zhang et al., 2025) accepts scripts, video, and pose and produces multi-speaker spatial dialog with dramatic prosody. MusicGen (Copet et al., 2023), Moûsai (Schneider et al., 2023), and (Evans

et al., 2024b) generate high-quality stereo music from text input.

Environmental Acoustic Modeling For higher realism and interactivity, research splits into two philosophies: holistic and compositional. Environmental acoustic modeling represents the holistic view. (Ratnarajah and Manocha, 2024) renders sound for a 3D scene with a graph neural network that encodes material and geometry. (Kim et al., 2019) estimates room geometry and acoustics from 360° images to synthesize scene-aware audio. Modular and zero-shot generation illustrates the compositional view. SEE-2-Sound (Dagli et al., 2024) breaks the visual-to-audio task into region recognition, 3D localization, mono generation, and spatialization. The modular design lets the system produce matching spatial audio for novel visual content and shows strong generalization.

4.3 Future Work

Spatial audio generation has made strong progress. The next challenge is to move beyond signal level reconstruction toward semantics driven generation. This also follows the from perception to cognition shift. Works such as ISDrama(Zhang et al., 2025) show the potential to generate dialogue style spatial audio from scripts and video that matches context and emotion. This suggests a move from simple spatialization to the creation of complex soundscapes with narrative logic and affect. Future research should deepen semantic control. A model should not only generate audio for a text like birds chirp on the left. It should also capture fine grained emotion and ambience. This requires richer semantics and better context understanding. It fits current trends in multimodal generation.

A critical but under explored direction is diversity and flexibility of output formats. Current models often support only a single output format and lack conversion across spatial representations. Future generators should be format agnostic and able to produce multiple spatial representations according to user instructions. Most works focus on input side fusion. Output side diversity and semantic generation remain open and challenging.

5 Dataset and Evaluation of Spatial Audio

5.1 Datasets

Spatial audio data exists in a variety of formats, each reflecting different characteristics and tailored to specific tasks. This section provides an in-depth analysis of existing spatial audio datasets, illustrating the diverse methods of data collection and processing, and explaining how these elements contribute to the understanding of spatial audio. Sources including real-world recordings, physics-based simulations, and web-crawled material are shown in Appendix E and Table 6.

5.1.1 Multi-Channel Audio Datasets

Multi-channel datasets are crucial for developing far-field speech interaction and scene analysis systems. Early corpora like REVERB Challenge (Kinoshita et al., 2016), DIRHA (Ravanelli et al., 2015), and Sweet-Home (Vacher et al., 2014) focus on speech enhancement and ASR in reverberant home environments. To support more precise spatial hearing research, datasets such as Voice-Home (Bertin et al., 2016), SECL-UMons (Brousmiche et al., 2020), and AVRI (Qian et al., 2022) provide detailed geometric annotations for localization and speaker tracking. Recent efforts capture dynamic and complex scenes, including pedestrian environments in the Wearable SELD dataset (Nagatomo et al., 2022) and diverse indoor/outdoor settings in the high-channel-count RealMAN dataset (Yang et al., 2024).

5.1.2 First-Order Ambisonics Datasets

First-Order Ambisonics (FOA) is a standard format for tasks requiring 3D acoustic information, with datasets collected via crawling, simulation, and real-world recording. Crawled datasets like YT-ALL (Morgado et al., 2018) and YT-360 (Morgado et al., 2020) provide large-scale, in-the-wild data for pre-training, while YT-AMBIGEN (Kim et al., 2025) improves alignment by filtering for camera metadata. Simulated datasets, including the TUT Sound Events series (Adavanne et al., 2018a) and DCASE2021 Task 3 (Politis et al., 2021), offer controlled benchmarks for SELD, whereas Spatial LibriSpeech (Sarabia et al., 2023) and SonicSet (Li et al., 2024a) spatialize large existing corpora. Scarce but highly realistic recorded datasets like REC-STREET (Morgado et al., 2018) and the STARSS series (Politis et al., 2022; Shimada et al., 2023) provide invaluable data for outdoor scenes and high-resolution SELD benchmarks.

5.1.3 Binaural Datasets

Binaural audio offers a perceptually plausible format for headphone-based immersion by directly mimicking human hearing. Real-world recordings

capture naturalistic scenes, from musical performances in FAIR-Play (Gao and Grauman, 2019) to challenging noisy conversations in EasyCom (Donley et al., 2021) and head-tracked dialogues in the dataset by Richard et al. (Richard et al., 2021). Simulated datasets like SimBinaural (Garg et al., 2023) enable large-scale, controllable data generation, while hybrid approaches like YouTube-Binaural (Garg et al., 2023) convert existing surround audio to a pseudo-binaural format. Recent efforts integrate richer multimodal and semantic information, with BEWO-1M (Sun et al., 2024) enabling text-guided generation and MRSDrama (Zhang et al., 2025) providing a unique corpus of expressive spatial speech for narrative tasks.

5.2 Objective Evaluation Metrics

5.2.1 Evaluation Metrics for Understanding

SELD Evaluation covers SED and DOA estimation. SED uses segment-based F-score and error rate (ER) (Mesaros et al., 2016). DOA uses two frame-wise metrics: DOA error, which measures the angular deviation between estimates and references, and frame recall, which measures the fraction of frames with the correct number of detected sources (Adavanne et al., 2018b). DOA error averages the assignment cost between reference DOAs $\mathbf{DOA}_R^t$ and estimated DOAs $\mathbf{DOA}_E^t$ based on the Hungarian algorithm.

Spatial Audio Separation Separation quality is measured with `mir_eval` metrics such as signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR)(Ye et al., 2024).

Joint Learning For audio–visual tasks, evaluation often uses binary classification metrics, such as audio–visual correspondence (AVC-Bin) and audio–visual spatial alignment (AVSA-Bin) (Morgado et al., 2020). Downstream tasks, such as semantic segmentation, use pixel accuracy and mean Intersection over Union (mIoU).

5.2.2 Evaluation Metrics for Generation

Monaural-to-Binaural Audio Generation Fidelity is evaluated with objective measures in the time domain (Wave L_2), spectral domain (Amplitude L_2, Phase L_2, multi-resolution STFT loss), and perceptual scores (PESQ, MOS) (Leng et al., 2022; Liu et al., 2022). The multi-resolution STFT loss (MRSTFT) combines spectral convergence $\mathcal{L}_{\text{SC}}$ and log-magnitude loss $\mathcal{L}_{\text{mag}}$.

End-to-End Binaural Audio Generation Evaluation focuses on key spatial cues. Objective metrics include mean absolute error (MAE) of interaural phase difference (IPD) and interaural level difference (ILD) (Zhang et al., 2025). Perceptual evaluation often measures cosine similarity between angle/distance embeddings from a pretrained model (e.g., SPATIAL-AST (Zheng et al., 2024)) and those from generated audio.

End-to-End FOA Generation Evaluation combines spatial accuracy, codec quality, and perceptual plausibility (Heydari et al., 2025). Spatial accuracy reports errors of azimuth (θ), elevation (ϕ), and distance (d), which are derived from the intensity vector of FOA channels. The overall spatial-angle error $\Delta_{\text{Spatial-Angle}}$ is also reported (Van Brummelen, 2012). Codec quality uses STFT and Mel distances. Plausibility uses Fréchet Audio Distance (FAD) and KL divergence. The CLAP score measures consistency between text prompts and generated audio.

Detailed formulas are presented in Appendix F.

5.3 Subjective Evaluation Metrics

Objective metrics give reproducible baselines, but human perception is the final standard, especially for immersion. Subjective tests collect listener feedback on timbre, spatial impression, realism, and overall immersion. These aspects are hard for signal level metrics to capture. Rigorous subjective evaluation is therefore essential. Spatial audio related subjective evaluation metrics including MOS test, MUSHRA (Series, 2014)and A/B and ABX (Boley and Lester, 2009).

6 Conclusion

This paper presents a comprehensive survey of the rapidly advancing spatial audio field covering foundational spatial audio input and output representations; the core research paradigms of understanding and generation; and the landscape of datasets and evaluation metrics.We hope this survey serves as a valuable resource for researchers, further guiding future work and fostering innovation in immersive audio technology.

Limitations

While this survey provides a broad overview of the algorithmic and data-centric aspects of spatial audio, its scope has certain limitations, leaving several important areas underexplored.

First, our review is heavily centered on software, models, and datasets, with only a cursory treatment of the specialized hardware that underpins the entire spatial audio pipeline. We do not offer a detailed analysis of different microphone array geometries (e.g., spherical, tetrahedral), the design of dedicated audio processors (DSPs) for real-time rendering, or the technologies behind head-tracking sensors (e.g., IMUs) and their integration into consumer devices. A deeper dive into these hardware components would be necessary for a complete picture of the field's engineering challenges.

Second, while we touch upon perceptual concepts like HRTF personalization and evaluation metrics like MOS, the survey does not delve deeply into the fundamentals of psychoacoustics and human spatial hearing. A dedicated discussion on the perceptual mechanisms that enable sound localization and immersion would provide crucial context for the engineering solutions presented. Similarly, our section on evaluation metrics focuses extensively on objective, formula-based measures but does not detail the methodologies of subjective listening tests (e.g., MUSHRA, A/B testing), which remain the gold standard for assessing the perceptual quality of spatial audio systems.

Ethical Considerations

As spatial audio matures and spreads, its ethical challenges grow and deserve careful study. The first concern is privacy. As noted in our draft, deploying multi-microphone arrays in private and public spaces for high-fidelity capture increases the risk of surveillance without consent. Spatial audio can record content and also infer speaker positions, movement paths, and even headcounts. Reconstructing physical scenes from intercepted audio becomes possible. This goes beyond traditional wiretapping and is a deeper privacy threat. It is therefore crucial to develop strong protections, such as on-device processing and differential privacy.

Rapid progress in spatial audio generation brings new risks, especially audio forgery and misinformation. Advanced models can mimic a person's voice and place it in a plausible virtual space, creating highly deceptive "spatial audio deepfakes." For example, an attacker could forge audio that sounds like a public figure speaking in a specific room. The spatial realism greatly boosts credibility and can be used to manipulate opinion, commit fraud, or harm reputations. This can erode trust in digital media. Detection methods and clear accountability frameworks are urgent needs.

We must also address bias and fairness. Core technologies such as personalized HRTF often rely on datasets measured on specific populations or on standardized head models. Lack of diversity in anthropometric traits can lead to unequal experiences across gender, ethnicity, or age. Some users may have poorer immersion and localization accuracy. Large models trained on web audio can also inherit and amplify social biases. The community should build inclusive datasets and perform fairness audits to ensure access and to avoid a new digital divide.

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant No.U24A20326.

References

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

Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. 2018b. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 1462–1466. IEEE.

Byeongjoo Ahn, Karren Yang, Brian Hamilton, Jonathan Sheaffer, Anurag Ranjan, Miguel Sarabia, Oncel Tuzel, and Jen-Hao Rick Chang. 2023. Novel-view acoustic synthesis from 3d reconstructed rooms. *arXiv preprint arXiv:2310.15130*.

Lior Arbel, Ishwarya Ananthabhotla, Zamir Ben-Hur, David Lou Alon, and Boaz Rafaely. 2024. On hrtf notch frequency prediction using anthropometric features and neural networks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 816–820. IEEE.

Jacob Benesty, Jingdong Chen, and Yiteng Huang. 2008. *Microphone array signal processing*. Springer.

Nancy Bertin, Ewen Camberlein, Emmanuel Vincent, Romain Lebarbenchon, Stéphane Peillon, Éric Lamandé, Sunit Sivasankaran, Frédéric Bimbot, Irina Illina, Ariane Tom, and 1 others. 2016. A french corpus for distant-microphone speech processing in real homes. In *Interspeech 2016*.

Michael J Bianco, Sharon Gannot, Efren Fernandez-Grande, and Peter Gerstoft. 2021. Semi-supervised source localization in reverberant environments with deep generative modeling. *IEEE Access*, 9:84956–84970.

Michael J Bianco, Sharon Gannot, and Peter Gerstoft. 2020. Semi-supervised source localization with deep generative modeling. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE.

Miguel Blanco Galindo, Philip Coleman, and Philip JB Jackson. 2020. Microphone array geometries for horizontal spatial audio object capture with beamforming. *Journal of the Audio Engineering Society*, 68(5):324–337.

Jon Boley and Michael Lester. 2009. Statistical analysis of abx results using signal detection theory. In *Audio engineering society convention*, volume 127.

Fabian Brinkmann, Alexander Lindau, and Stefan Weinzierl. 2017. On the authenticity of individual dynamic binaural synthesis. *The Journal of the Acoustical Society of America*, 142(4):1784–1795.

James Broderick, Jim Duggan, and Sam Redfern. 2018. The importance of spatial audio in modern games and virtual environments. In *2018 IEEE games, entertainment, media conference (GEM)*, pages 1–9. IEEE.

Mathilde Brousmiche, Jean Rouat, and Stéphane Dupont. 2020. Secl-umons database for sound event classification and localization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 756–760. IEEE.

C Phillip Brown and Richard O Duda. 1998. A structural model for binaural sound synthesis. *IEEE transactions on speech and audio processing*, 6(5):476–488.

Nicholas J Bryan. 2020. Impulse response data augmentation and deep neural networks for blind room acoustic parameter estimation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yin Cao, Turab Iqbal, Qiuqiang Kong, Fengyan An, Wenwu Wang, and Mark D Plumbley. 2021. An improved event-independent network for polyphonic sound event localization and detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 885–889. IEEE.

Yin Cao, Qiuqiang Kong, Turab Iqbal, Fengyan An, Wenwu Wang, and Mark D. Plumbley. 2019. Polyphonic sound event detection and localization using a two-stage strategy. *CoRR*, abs/1905.00268.

Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. 2022. Visual acoustic matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18858–18868.

Guodong Chen, Sizhe Wang, Jacob Chakareski, Dimitrios Koutsonikolas, and Mallesham Dasari. 2025. Spatial video streaming on apple vision pro xr headset. In *Proceedings of the 26th International Workshop on Mobile Computing Systems and Applications*, pages 115–120.

Xingyu Chen, Fei Ma, Yile Zhang, Amy Bastine, and Prasanga N Samarasinghe. 2023. Head-related transfer function interpolation with a spherical cnn. *arXiv preprint arXiv:2309.08290*.

Jozef Coldenhoff, Andrew Harper, Paul Kendrick, Tijana Stojkovic, and Milos Cernak. 2024. Multichannel mosra: Mean opinion score and room acoustics estimation using simulated data and a teacher model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 381–385. IEEE.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36:47704–47720.

Rishit Dagli, Shivesh Prakash, Robert Wu, and Houman Khosravani. 2024. See-2-sound: Zero-shot spatial environment-to-spatial sound. *arXiv preprint arXiv:2406.06612*.

Antoine Deleforge and Radu Horaud. 2012. The cocktail party robot: Sound source separation and localisation with an active binaural head. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 431–438.

Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. 2024. Learning spatially-aware language and audio embeddings. *Advances in Neural Information Processing Systems*, 37:33505–33537.

Diego Di Carlo, Pinchas Tandeitnik, Cédric Foy, Antoine Deleforge, Nancy Bertin, and Sharon Gannot. 2021. dechorate: a calibrated room impulse response database for echo-aware signal processing. *arXiv preprint arXiv:2104.13168*.

Jacob Donley, Vladimir Tourbabin, Jung-Suk Lee, Mark Broyles, Hao Jiang, Jie Shen, Maja Pantic, Vamsi Krishna Ithapu, and Ravish Mehra. 2021. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174*.

Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons. 2024a. Fast timing-conditioned latent audio diffusion. In *Forty-first International Conference on Machine Learning*.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2024b. Long-form music generation with latent diffusion. *arXiv preprint arXiv:2404.10301*.

George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *Acm Sigkdd Explorations Newsletter*, 12(1):49–57.

Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. 2019. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7053–7062.

Ruohan Gao and Kristen Grauman. 2019. 2.5 d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333.

Guillermo García-Barrios, Daniel Aleksander Krause, Archontis Politis, Annamaria Mesaros, Juana M Gutiérrez-Arriola, and Rubén Fraile. 2022. Binaural source localization using deep learning and head rotation information. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 36–40. IEEE.

Rishabh Garg, Ruohan Gao, and Kristen Grauman. 2021. Geometry-aware multi-task learning for binaural audio generation from video. *arXiv preprint arXiv:2111.10882*.

Rishabh Garg, Ruohan Gao, and Kristen Grauman. 2023. Visually-guided audio spatialization in video with geometry-aware multi-task learning. *International Journal of Computer Vision*, 131(10):2723–2737.

François Grondin, Jean-Samuel Lauzon, Simon Michaud, Mirco Ravanelli, and François Michaud. 2020. Bird: Big impulse response dataset. *arXiv preprint arXiv:2010.09930*.

Cong Han, Yi Luo, and Nima Mesgarani. 2020. Real-time binaural speech separation with preserved spatial cues. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6404–6408. IEEE.

Mojtaba Heydari, Mehrez Souden, Bruno Conejo, and Joshua Atkins. 2025. Immersediffusion: A generative spatial audio latent diffusion model. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Aidan OT Hogg, Mads Jenkins, He Liu, Isaac Squires, Samuel J Cooper, and Lorenzo Picinali. 2024. Hrtf upsampling with a generative adversarial network using a gnomonic equiangular projection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Wen Chin Huang, Dejan Markovic, Alexander Richard, Israel Dejene Gebru, and Anjali Menon. 2022. End-to-end binaural speech synthesis. *arXiv preprint arXiv:2207.03697*.

Yankun Huang, Xihong Wu, and Tianshu Qu. 2020. A time-domain unsupervised learning based sound source localization method. In *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, pages 26–32. IEEE.

E ISO. 2009. 3382-1, 2009,"acoustics—measurement of room acoustic parameters—part 1: Performance spaces,". *International Organization for Standardization, Brussels, Belgium*, 69.

Yuki Ito, Tomohiko Nakamura, Shoichi Koyama, and Hiroshi Saruwatari. 2022. Head-related transfer function interpolation from spatially sparse measurements using autoencoder with source position conditioning. In *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5. IEEE.

Marco Jeub, Magnus Schafer, and Peter Vary. 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *2009 16th international conference on digital signal processing*, pages 1–5. IEEE.

Ziran Jiang, Jinqiu Sang, Chengshi Zheng, Andong Li, and Xiaodong Li. 2023. Modeling individual head-related transfer functions from sparse measurements using a convolutional neural network. *The Journal of the Acoustical Society of America*, 153(1):248–259.

Hansung Kim, Luca Remaggi, Philip JB Jackson, and Adrian Hilton. 2019. Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 120–126. IEEE.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. 2025. Visage: Video-to-spatial audio generation. In *The Thirteenth International Conference on Learning Representations*.

Jin Sob Kim, Hyun Joon Park, Wooseok Shin, and Sung Won Han. 2023. Ad-yolo: You look only once in training multiple sound event localization and detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Keisuke Kinoshita, Marc Delcroix, Sharon Gannot, Emanuël A P. Habets, Reinhold Haeb-Umbach, Walter Kellermann, Volker Leutnant, Roland Maas, Tomohiro Nakatani, Bhiksha Raj, and 1 others. 2016. A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, 2016:1–19.

Shoichi Koyama, Tomoya Nishida, Keisuke Kimura, Takumi Abe, Natsuki Ueno, and Jesper Brunnström. 2021. Meshrir: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods. In *2021 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, pages 1–5. IEEE.

Daniel Aleksander Krause, Guillermo García-Barrios, Archontis Politis, and Annamaria Mesaros. 2023. Binaural sound source distance estimation and localization for a moving listener. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Sheng Kuang, Jie Shi, Kiki van der Heijden, and Siamak Mehrkanoon. 2022. Bast: Binaural audio spectrogram transformer for binaural sound localization. *arXiv preprint arXiv:2207.03927*.

Saksham Singh Kushwaha, Jianbo Ma, Mark RP Thomas, Yapeng Tian, and Avery Bruni. 2025. Diffsage: End-to-end spatial audio generation using diffusion models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ben Lee, Tomasz Rudzki, Jan Skoglund, and Gavin Kearney. 2023a. Context-based evaluation of the opus audio codec for spatial audio content in virtual reality. *Journal of the Audio Engineering Society*, 71:145–154.

Jin Woo Lee and Kyogu Lee. 2023. Neural fourier shift for binaural speech rendering. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jin Woo Lee, Sungho Lee, and Kyogu Lee. 2023b. Global hrtf interpolation via learned affine transformation of hyper-conditioned features. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yichong Leng, Zehua Chen, Junliang Guo, Haohe Liu, Jiawei Chen, Xu Tan, Danilo Mandic, Lei He, Xiangyang Li, Tao Qin, and 1 others. 2022. Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. *Advances in Neural Information Processing Systems*, 35:23689–23700.

Kai Li, Wendi Sang, Chang Zeng, Runxuan Yang, Guo Chen, and Xiaolin Hu. 2024a. Sonicsim: A customizable simulation platform for speech processing in moving sound source scenarios. *arXiv preprint arXiv:2410.01481*.

Sijia Li, Shiguang Liu, and Dinesh Manocha. 2021. Binaural audio generation via multi-task learning. *ACM Transactions on Graphics (TOG)*, 40(6):1–13.

Zhaojian Li, Bin Zhao, and Yuan Yuan. 2024b. Cross-modal generative model for visual-guided binaural stereo generation. *Knowledge-Based Systems*, 296:111814.

Zhaojian Li, Bin Zhao, and Yuan Yuan. 2024c. Cyclic learning for binaural audio generation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26669–26678.

Susan Liang, Chao Huang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. 2023. Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *Advances in Neural Information Processing Systems*, 36:37472–37490.

Yan-Bo Lin and Yu-Chiang Frank Wang. 2021. Exploiting audio-visual consistency with partial supervision for spatial audio generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2056–2063.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Huadai Liu, Tianyi Luo, Qikai Jiang, Kaicheng Luo, Peiwen Sun, Jialei Wan, Rongjie Huang, Qian Chen, Wen Wang, Xiangtai Li, and 1 others. 2025a. Omniaudio: Generating spatial audio from 360-degree video. *arXiv preprint arXiv:2504.14906*.

Jinglin Liu, Zhenhui Ye, Qian Chen, Siqi Zheng, Wen Wang, Qinglin Zhang, and Zhou Zhao. 2022. Dopplerbas: Binaural audio synthesis addressing doppler effect. *arXiv preprint arXiv:2212.07000*.

Miao Liu, Jing Wang, Xinyuan Qian, and Xiang Xie. 2024. Visually guided binaural audio generation with cross-modal consistency. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7980–7984. IEEE.

Xiaojing Liu, Ogulcan Gurelli, Yan Wang, and Joshua Reiss. 2025b. Visual-based spatial audio generation system for multi-speaker environments. *arXiv preprint arXiv:2502.07538*.

Tim Lübeck, Johannes M Arend, and Christoph Pörschmann. 2022. Binaural reproduction of dummy head and spherical microphone array data—a perceptual study on the minimum required spatial resolution. *The Journal of the Acoustical Society of America*, 151(1):467–483.

Andrew Luo, Yilun Du, Michael Tarr, Josh Tenenbaum, Antonio Torralba, and Chuang Gan. 2022. Learning neural acoustic fields. *Advances in Neural Information Processing Systems*, 35:3165–3177.

Fei Ma, Thushara D Abhayapala, Prasanga N Samarasinghe, and Xingyu Chen. 2023. Spatial upsampling of head-related transfer functions using a physics-informed neural network. *arXiv preprint arXiv:2307.14650*.

Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. 2022. Few-shot audio-visual learning of environment acoustics. *Advances in Neural Information Processing Systems*, 35:2522–2536.

David G Malham and Anthony Myatt. 1995. 3-d sound spatialization using ambisonic techniques. *Computer music journal*, 19(4):58–70.

Nils Marggraf-Turley, Michael Lovedee-Turner, and Enzo De Sena. 2024. Hrtf recommendation based on the predicted binaural colouration model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1106–1110. IEEE.

Yoshiki Masuyama, Gordon Wichern, François G Germain, Zexu Pan, Sameer Khurana, Chiori Hori, and Jonathan Le Roux. 2024. Niirf: Neural iir filter field for hrtf upsampling and personalization. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1016–1020. IEEE.

Tobias May, Steven Van De Par, and Armin Kohlrausch. 2010. A probabilistic model for robust localization based on a binaural auditory front-end. *IEEE Transactions on audio, speech, and language processing*, 19(1):1–13.

Thomas McKenzie, Leo McCormack, and Christoph Hold. 2021a. Dataset of spatial room impulse responses in a variable acoustics room for six degrees-of-freedom rendering and analysis. *arXiv preprint arXiv:2111.11882*.

Thomas McKenzie, Sebastian J Schlecht, and Ville Pulkki. 2021b. Acoustic analysis and dataset of transitions between coupled rooms. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 481–485. IEEE.

Annamaria Mesaros, Toni Heittola, Aleksandr Diment, Benjamin Elizalde, Ankit Shah, Emmanuel Vincent, Bhiksha Raj, and Tuomas Virtanen. 2017. Dcase 2017 challenge setup: Tasks, datasets and baseline system. In *DCASE 2017-workshop on detection and classification of acoustic scenes and events*.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162.

Christina Mittag, M Böhme, S Werner, and S Klein. 2017. Dataset of binaural room impulse responses at multiple recording positions, source positions, and orientations in a real room. In *in Proc. of the 43rd annual convention for acoustics, DAGA, Germany*.

Brian CJ Moore. 2012. *An introduction to the psychology of hearing*. Brill.

Pedro Morgado, Yi Li, and Nuno Nvasconcelos. 2020. Learning representations from audio-visual spatial alignment. *Advances in Neural Information Processing Systems*, 33:4733–4744.

Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. 2018. Self-supervised generation of spatial audio for 360 video. *Advances in neural information processing systems*, 31.

David Murphy and Flaithrí Neff. 2011. Spatial sound for computer games and virtual reality. In *Game sound technology and player interaction: Concepts and developments*, pages 287–312. IGI Global Scientific Publishing.

Kento Nagatomo, Masahiro Yasuda, Kohei Yatabe, Shoichiro Saito, and Yasuhiro Oikawa. 2022. Wearable seld dataset: Dataset for sound event localization and detection using wearable devices around head. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 156–160. IEEE.

Javier Naranjo-Alcazar, Sergi Perez-Castanos, Jose Ferrandis, Pedro Zuccarello, and Maximo Cobos. 2020. Sound event localization and detection using squeeze-excitation residual cnns. *arXiv preprint arXiv:2006.14436*.

Thi Ngoc Tho Nguyen, Karn N Watcharasupat, Ngoc Khanh Nguyen, Douglas L Jones, and Woon-Seng Gan. 2022. Salsa: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1749–1762.

Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. 2016. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664.

Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. 2022. Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3347–3356.

Despoina Pavlidi, Symeon Delikaris-Manias, Ville Pulkki, and Athanasios Mouchtaris. 2015. 3d localization of multiple sound sources with intensity vector estimates in single source zones. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1556–1560. IEEE.

Sandra Poeschl, Konstantin Wall, and Nicola Doering. 2013. Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence. In *2013 IEEE Virtual Reality (VR)*, pages 129–130. IEEE.

Archontis Politis, Sharath Adavanne, Daniel Krause, Antoine Deleforge, Prerak Srivastava, and Tuomas Virtanen. 2021. A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection. *arXiv preprint arXiv:2106.06999*.

Archontis Politis, Sharath Adavanne, and Tuomas Virtanen. 2020. A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection. *arXiv preprint arXiv:2006.01919*.

Archontis Politis, Kazuki Shimada, Parthasaarathy Sudarsanam, Sharath Adavanne, Daniel Krause, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, Yuki Mitsufuji, and Tuomas Virtanen. 2022. Starss22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *arXiv preprint arXiv:2206.01948*.

Xinyuan Qian, Zhengdong Wang, Jiadong Wang, Guohui Guan, and Haizhou Li. 2022. Audio-visual cross-attention network for robotic speaker tracking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:550–562.

Kranthi Kumar Rachavarapu, Vignesh Sundaresha, AN Rajagopalan, and 1 others. 2021. Localize to binauralize: Audio spatialization from visual sound source localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1930–1939.

Aakanksha Rana, Cagri Ozcinar, and Aljosa Smolic. 2019. Towards generating ambisonics using audio-visual cue for virtual reality. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2012–2016. IEEE.

Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, and Dinesh Manocha. 2024. Av-rir: Audio-visual room impulse response estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27164–27175.

Anton Ratnarajah and Dinesh Manocha. 2024. Listen2scene: Interactive material-aware binaural sound propagation for reconstructed 3d scenes. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pages 254–264. IEEE.

Anton Ratnarajah, Zhenyu Tang, Rohith Aralikatti, and Dinesh Manocha. 2022. Mesh2ir: Neural acoustic impulse response generator for complex 3d scenes. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 924–933.

Mirco Ravanelli, Luca Cristoforetti, Roberto Gretter, Marco Pellin, Alessandro Sosi, and Maurizio Omologo. 2015. The dirha-english corpus and related tasks for distant-speech recognition in domestic environments. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 275–282. IEEE.

Alexander Richard, Dejan Markovic, Israel D Gebru, Steven Krenn, Gladstone Alexander Butler, Fernando Torre, and Yaser Sheikh. 2021. Neural synthesis of binaural speech from mono audio. In *International Conference on Learning Representations*.

Iran R Roman, Christopher Ick, Sivan Ding, Adrian S Roman, Brian McFee, and Juan P Bello. 2024. Spatial scaper: a library to simulate and augment soundscapes for sound event localization and detection in realistic rooms. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1221–1225. IEEE.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer.

Orlem Santos, Karen Rosero, Bruno Masiero, and Roberto de Alencar Lotufo. 2024. w2v-seld: A sound event localization and detection framework for self-supervised spatial audio pre-training. *IEEE Access*.

Miguel Sarabia, Elena Menyaylenko, Alessandro Toso, Skyler Seto, Zakaria Aldeneh, Shadi Pirhosseinloo, Luca Zappella, Barry-John Theobald, Nicholas Apostoloff, and Jonathan Sheaffer. 2023. Spatial librispeech: An augmented dataset for spatial audio learning. *arXiv preprint arXiv:2308.09514*.

Lauri Savioja and U Peter Svensson. 2015. Overview of geometrical room acoustic modeling techniques. *The Journal of the Acoustical Society of America*, 138(2):708–730.

Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. 2023. Mo\^ usai: Text-to-music generation with long-context latent diffusion. *arXiv preprint arXiv:2301.11757*.

B Series. 2014. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly*, 2.

Kazuki Shimada, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, and Yuki Mitsufuji. 2021. Accdoa: Activity-coupled cartesian direction of arrival representation for sound event localization and detection. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 915–919. IEEE.

Kazuki Shimada, Yuichiro Koyama, Shusuke Takahashi, Naoya Takahashi, Emiru Tsunoo, and Yuki Mitsufuji. 2022. Multi-accdoa: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 316–320. IEEE.

Kazuki Shimada, Archontis Politis, Parthasaarathy Sudarsanam, Daniel A Krause, Kengo Uchida, Sharath Adavanne, Aapo Hakala, Yuichiro Koyama, Naoya Takahashi, Shusuke Takahashi, and 1 others. 2023. Starss23: An audio-visual dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events. *Advances in neural information processing systems*, 36:72931–72957.

Prerak Srivastava, Antoine Deleforge, and Emmanuel Vincent. 2021. Blind room parameter estimation using multiple multichannel speech recordings. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 226–230. IEEE.

Peiwen Sun, Sitong Cheng, Xiangtai Li, Zhen Ye, Huadai Liu, Honggang Zhang, Wei Xue, and Yike Guo. 2024. Both ears wide open: Towards language-driven spatial audio generation. *arXiv preprint arXiv:2410.10676*.

Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. 2023. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6420–6429.

Christian Templin, Yanda Zhu, and Hao Wang. 2025. Sonicmotion: Dynamic spatial audio soundscapes with latent diffusion models. *arXiv preprint arXiv:2507.07318*.

Etienne Thuillier, Craig T Jin, and Vesa Välimäki. 2024. Hrtf interpolation using a spherical neural process meta-learner. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1790–1802.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263.

Michel Vacher, Benjamin Lecouteux, Pedro Chahuara, François Portet, Brigitte Meillon, and Nicolas Bonnefond. 2014. The sweet-home speech and multimodal corpus for home automation interaction. In *The 9th edition of the Language Resources and Evaluation Conference (LREC)*, pages 4499–4506.

Glen Van Brummelen. 2012. *Heavenly mathematics: The forgotten art of spherical trigonometry*. Princeton University Press.

Rejin Varghese and M Sambath. 2024. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6. IEEE.

Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey. 2018. Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 1–5. IEEE.

Zhong-Qiu Wang and DeLiang Wang. 2018. Combining spectral and spatial features for deep learning based blind speaker separation. *IEEE/ACM Transactions on audio, speech, and language processing*, 27(2):457–468.

Michaela Warnecke, Sharon Jamison, Sebastian Prepelita, Paul Calamia, and Vamsi Krishna Ithapu. 2022. Hrtf personalization based on ear morphology. In *Audio Engineering Society Conference: 2022 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society.

Ron J Weiss, Michael I Mandel, and Daniel PW Ellis. 2009. Source separation based on binaural cues and source model constraints.

Yifan Wu, Roshan Ayyalasomayajula, Michael J Bianco, Dinesh Bharadia, and Peter Gerstoft. 2021. Sslide: Sound source localization for indoors based on deep learning. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4680–4684. IEEE.

Lauri Wuolio and Elina Moreira Kares. 2023. On the potential of spatial audio in enhancing virtual user experiences.

Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. 2021. Visually informed binaural audio generation without binaural audios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494.

Bing Yang, Changsheng Quan, Yabo Wang, Pengyu Wang, Yujie Yang, Ying Fang, Nian Shao, Hui Bu, Xin Xu, and Xiaofei Li. 2024. Realman: A real-recorded and annotated microphone array dataset for dynamic speech enhancement and localization. *Advances in Neural Information Processing Systems*, 37:105997–106019.

Haici Yang, Sanna Wager, Spencer Russell, Mike Luo, Minje Kim, and Wontak Kim. 2022. Upmixing via style transfer: a variational autoencoder for disentangling spatial images and musical content. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 426–430. IEEE.

Karren Yang, Bryan Russell, and Justin Salamon. 2020. Telling left from right: Learning spatial correspondence of sight and sound. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9932–9941.

Qiang Yang and Yuanqing Zheng. 2022. Deepear: Sound localization with binaural microphones. *IEEE Transactions on Mobile Computing*, 23(1):359–375.

Masahiro Yasuda, Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, and Keisuke Imoto. 2020. Sound event localization based on sound intensity vector refined by dnn-based denoising and source separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 651–655. IEEE.

Yuxin Ye, Wenming Yang, and Yapeng Tian. 2024. Lavss: Location-guided audio-visual spatial audio separation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5508–5519.

Yongyi Zang, Yifan Wang, and Minglun Lee. 2024. Ambisonizer: Neural upmixing as spherical harmonics generation. *arXiv preprint arXiv:2405.13428*.

Wen Zhang and Jie Shao. 2021. Multi-attention audio-visual fusion network for audio spatialization. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 394–401.

Xueliang Zhang and DeLiang Wang. 2017. Deep learning based binaural speech separation in reverberant environments. *IEEE/ACM transactions on audio, speech, and language processing*, 25(5):1075–1084.

You Zhang, Yuxiang Wang, and Zhiyao Duan. 2023. Hrtf field: Unifying measured hrtf magnitude representation with neural fields. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yu Zhang, Wenxiang Guo, Changhao Pan, Zhiyuan Zhu, Tao Jin, and Zhou Zhao. 2025. Isdrama: Immersive spatial drama generation through multimodal prompting. *arXiv preprint arXiv:2504.20630*.

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)*, pages 570–586.

Lei Zhao, Sizhou Chen, Linfeng Feng, Xiao-Lei Zhang, and Xuelong Li. 2025. Dualspec: Text-to-spatial-audio generation via dual-spectrogram guided diffusion model. *arXiv preprint arXiv:2502.18952*.

Manlin Zhao, Zhichao Sheng, and Yong Fang. 2022. Magnitude modeling of personalized hrtf based on ear images and anthropometric measurements. *Applied Sciences*, 12(16):8155.

Tao Zheng, Sunny Verma, and Wei Liu. 2022. Interpretable binaural ratio for visually guided binaural audio generation. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. 2024. Bat: Learning to reason about spatial sounds with large language models. *arXiv preprint arXiv:2402.01591*.

Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. 2020. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 52–69. Springer.

Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. 2018. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3550–3558.

Katerina Zmolikova, Marc Delcroix, Lukáš Burget, Tomohiro Nakatani, and Jan Honza Černocky. 2021. Integration of variational autoencoder and spatial clustering for adaptive multi-channel neural speech separation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 889–896. IEEE.

Franz Zotter and Matthias Frank. 2019. *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer Nature.

Appendix

A Extended Representation Discussions

A.1 Input Representationsn

Figure 2 shows the input modalities for spatial audio tasks, natural language, spatial position, visual information, and monaural audio, each offers a unique perspective for the system to perceive, interpret, or generate soundscapes. While they can be used independently, their true potential is often realized through synergistic multimodal combinations. The choice of input representation is not merely a technical decision but a fundamental architectural one that dictates the system's capabilities, complexity, and the nature of its interaction with the user or environment. This section will comparatively analyze these input paradigms, examining their intrinsic properties, task suitability, and the emerging trends in their combined application.

As shown in Table 1, these input representations exhibit a core trade-off between the level of abstraction and control precision. Natural language and visual information reside at the highest level of abstraction. They are intuitive for humans and well-suited for high-level scene description or content querying. However, this intuitiveness introduces challenges of lower control precision and semantic ambiguity, necessitating complex models to bridge the gap between semantics and machine-processable signals.

Conversely, spatial position coordinates offer the highest control precision, making them ideal for defining precise source trajectories or serving as ground truth for evaluation. However, they lack semantic context, and manually specifying complex scenes is a tedious process. Monaural audio plays a unique role. Positioned at a low level of abstraction, it does not directly provide spatial control. Instead, it serves as the foundational acoustic content for generation tasks, providing core acoustic features such as timbre and pitch. It acts as raw material that other modalities spatialize.

We observed that some multimodal works compare different input forms, for example the multimodal input in ISDrama(Zhang et al., 2025). As shown in Table 2, it suggests that precise geometric coordinates give the best scores. Video inputs are slightly worse. The model can learn relative spatial cues from video but lacks the precision of geometry. Text inputs perform the worst. They provide only coarse spatial hints and lag behind in ANG and DIS cosine similarity, which indicates less accurate angle and distance estimation.

Therefore, the selection of an input representation is fundamentally a trade-off between the intuitive, abstract control preferred by humans and the precise, geometric data required by machines, a choice contingent on the specific requirements of the task.

Abstract intent vs. geometric precision A fundamental trade-off exists among the different input representations: the opposition between the level of abstraction in control and its precision. Natural language and visual information represent the pinnacle of abstract, human-centric control. Natural language provides an intuitive way to specify semantic content (e.g., "a bird is chirping") and relational spatial attributes ("on the left"). Similarly, visual information from images or videos offers rich spatial and semantic context. These inputs describe what exists in a scene and how its components are related, which aligns closely with human perception.

However, this intuitiveness comes at the cost of reduced precision. The system must infer precise physical parameters from abstract descriptions. The BAT model (Zheng et al., 2024) exemplifies this challenge, utilizing a large language model to interpret complex natural language queries regarding "sound event detection, direction and distance estimation, and spatial reasoning". This highlights a critical point: high-level abstract inputs require a sophisticated, AI-based interpretation layer to translate human intent into machine-executable instructions.

In contrast, spatial position data provides the highest degree of precision. Cartesian or spherical coordinates offer direct and unambiguous guidance for placing sound sources. This makes it indispensable for tasks requiring absolute accuracy, such as providing ground truth for training and evaluating sound localization models, or simulating precise physical phenomena like the Doppler effect by incorporating velocity vectors (Liu et al., 2022; Zhang et al., 2025). The inherent trade-off is that this representation lacks semantic context and is non-intuitive and tedious for manually specifying complex acoustic scenes.

Monaural audio as the acoustic substrate Unlike other inputs that primarily define where a sound is, monaural audio defines what the sound itself is. It constitutes the "foundational acoustic

Attribute	Natural Language	Spatial Position	Visual Information	Monaural Audio
Primary Info	Semantic, relational, implicit spatial	Explicit spatial, dynamic	Semantic, spatial, dynamic	Acoustic (timbre, pitch, content)
Control Precision	Low	Very high	High	N/A
Abstraction Level	High	Low	High	Low
Interpretability	Indirect	Direct	Indirect	Indirect
Key Challenges	Ambiguity; semantic–signal gap	No semantics; tedious authoring	Ambiguity; occlusion; compute cost	Lack of spatial cues

Table 1: Comparative analysis of spatial audio *input* representations.

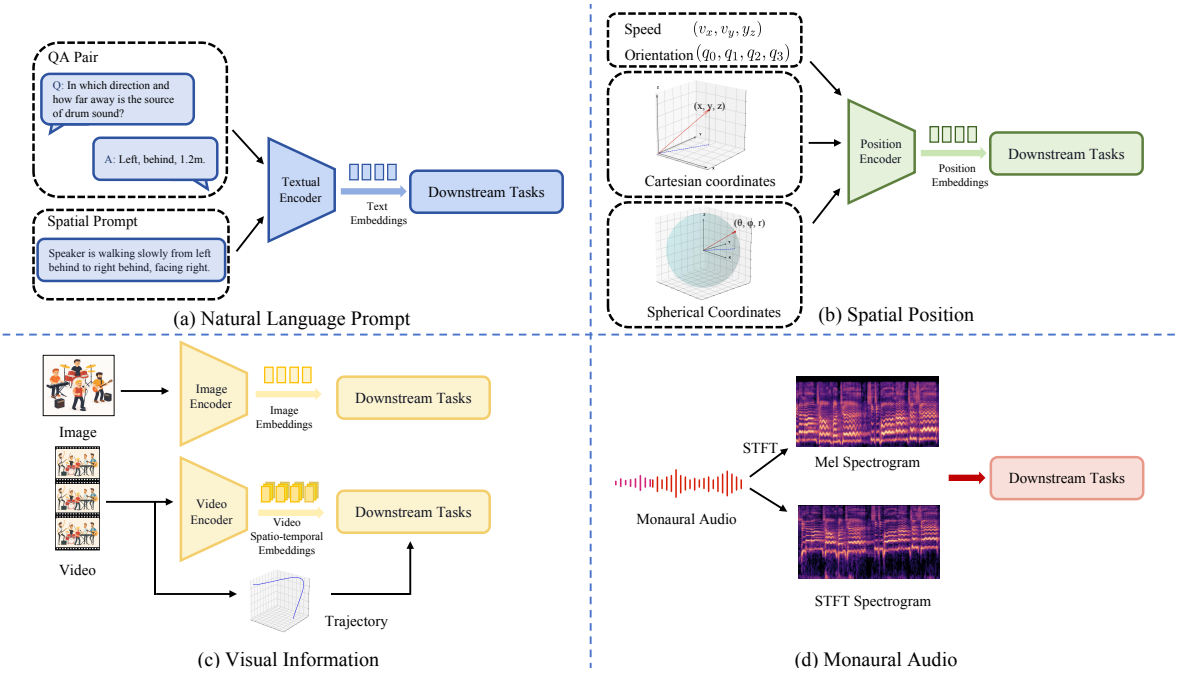

Figure 2: An overview of input representations of spatial audio and their primary processing methods.

content" for many spatial audio tasks, providing core acoustic characteristics such as timbre of a specific instrument or the phonetic features of speech. Therefore, monaural audio plays a unique role in the ecosystem of input representations.

Many advanced generative systems follow a two-stage principle: first, a source model (such as AudioGen (Kreuk et al., 2022) or AudioLDM (Liu et al., 2023)) generates a monaural audio stream; then, this stream is spatialized or upmixed into a multichannel or binaural format under the guidance of other input modalities, such as visual or positional data. This architecture clearly separates the problem of content generation from that of spatial rendering, enabling modular and flexible system design. Consequently, monaural audio is not an alternative option parallel to other input forms, but rather the fundamental substrate upon which they act.

Multimodal synergy The most powerful spatial audio systems are increasingly moving towards multimodality, creating comprehensive control schemes by combining the strengths of different input types to overcome the limitations of any single modality. The synergy between vision and audio is particularly potent. In audio-visual source separation tasks, the visual presence of an object (e.g., a speaking person) provides a strong, albeit implicit, cue for isolating its corresponding sound from a noisy mixture. In generation tasks, visual information can guide the spatialization process; for example, a U-Net architecture can take a monaural input and, guided by a video, render a spatially correct binaural or stereo output. The audio-visual matching task is considered crucial, highlighting the deeply learned correspondences between these modalities.

Similarly, adding explicit spatial position data

ISDrama Input Format	IPD MAE (↓)	ILD MAE (↓)	ANG COS (↑)	DIS COS (↑)
geometric	0.008	0.046	0.51	0.75
video	0.009	0.051	0.48	0.73
textual	0.011	0.055	0.43	0.68

Table 2: Quantitative Comparison of input modality, the data is from ISDrama(Zhang et al., 2025).

(such as source orientation and velocity) to a monaural audio stream allows for the simulation of highly realistic dynamic effects, like the Doppler shift , elevating realism to a level unattainable with static spatialization.

A.2 Output Representations

Table 3 presents a comparative analysis of the three primary spatial audio output representations. Each paradigm possesses unique advantages and limitations, making it suitable for different application scenarios and user requirements.

Playback system dependency and scalability are key to understanding the evolution of these three paradigms. Channel-based formats exhibit very high system dependency but poor scalability. This is because their audio mix is baked-in for a specific, standardized loudspeaker layout (e.g., 5.1 surround sound). Any playback system that deviates from this layout will degrade the intended spatial effect. In contrast, object-based formats feature low dependency and excellent scalability. They achieve this by decoupling the audio content from its metadata, which allows the playback device to render the audio in real-time according to its own arbitrary loudspeaker configuration. Consequently, a single master file can be adapted to any system. Scene-based formats occupy a middle ground. Their high dependency stems from the requirement for numerous loudspeakers and complex processing systems to physically reconstruct the sound field. Their moderate scalability is demonstrated by the ability to improve performance by increasing the system order (e.g., Higher-Order Ambisonics), though this significantly increases system cost and complexity.

Freedom of listening position and playback-end complexity are directly related to user experience and implementation cost. Channel-based formats confine the listener to a narrow sweet spot, but their playback-end complexity is low, requiring only simple channel-to-loudspeaker mapping. Scene-based formats offer high freedom, allowing listeners to move freely within a designated area. However, this comes at the cost of very high playback-end complexity, which involves real-time decoding and substantial signal processing. Object-based formats provide moderate freedom of movement (depending on the rendering system). Their moderate to high playback-end complexity arises from the need for a real-time rendering engine to process metadata and dynamically generate the mix.

Overall, these three paradigms are not mutually exclusive; rather, each has its optimal application domain. Channel-based technology retains its place in traditional media due to its simplicity and broad compatibility. Scene-based techniques offer irreplaceable advantages in applications requiring high physical fidelity and large-scale public experiences. Meanwhile, object-based technology, with its unparalleled flexibility and interactivity, has become the core driver for next-generation immersive media, such as VR/AR, gaming, and streaming. Understanding their fundamental differences is crucial for selecting and implementing the most appropriate spatial audio solution.

B Understanding Details

For SELD tasks, the metrics are relatively standardized. Competitions and datasets like DCASE challenges and TAU-NIGENS datasets provide evaluation metrics to evaluate the models' performance.

The diversity of datasets itself in table 4 reflects the lack of a unified benchmark which evaluate both objective and subjective quality of spatial audio.

C Generation Details

This section provides a detailed description of the input and output formats for the generative models summarized in Table 5. These formats represent the diverse ways in which spatial audio systems are controlled and the types of immersive experiences they can produce.

Spatial audio generation has evolved from two-stage upmixing approaches to fully end-to-end synthesis, driven by increasingly powerful deep learning architectures. Early and still prevalent methods,

Attribute	Channel-Based	Scene-Based	Object-Based
Freedom of Listening Position	Limited	High	Moderate
Playback System Dependency	Very high	High	Low
Scalability	Low	Moderate	Excellent
Playback-End Complexity	Low	High	Moderate
Common Formats	Stereo; 5.1/7.1 surround	Ambisonics; wave-field synthesis (WFS)	Dolby Atmos; DTS:X; MPEG-H 3D Audio

Table 3: Comparative analysis of spatial audio *output* representations.

Model	ER_{20}	F_{20}	LE_{CD}	LR_{CD}	ε_{SELD}	Dataset
CRNN-based SELD (Adavanne et al., 2018b)	0.04	97.7%	3.4°	99.4%	–	TUT Sound Events 2018
Naranjo-Alcazar et al. 2020	0.68	42.3%	22.5°	65.1%	–	DCASE2020
ACCDOA (Shimada et al., 2021)	0.43	74.2%	9.6°	77.5%	–	DCASE2020
Cao et al. 2021	0.233	83.2%	6.8°	86.1%	–	TAU Spatial Sound Events 2020
SALSA (Nguyen et al., 2022)	0.408	71.5%	12.6°	72.8%	0.259	TAU Spatial Sound Events 2019
Multi-ACCDOA (Shimada et al., 2022)	0.596	55.3%	18.4°	64.4%	0.375	DCASE2021
AD-YOLO (Kim et al., 2023)	0.4818	61.27%	8.48°	69.82%	0.3048	DCASE2022
w2v-SELD (Santos et al., 2024)	0.096	94.66%	4.67°	93.05%	0.061	TAU-2019

Table 4: Comparison of SELD models. ER_{20}: error rate at a 20° collar; F_{20}: F-score at a 20° collar; LE_{CD}: localization error; LR_{CD}: localization recall; ε_{SELD}: SELD score.

often based on CNNs like U-Net, focus on spatializing existing audio. These models typically take a monaural audio track and visual information from an image or video as input, and output a corresponding binaural or multi-channel audio signal, as seen in pioneering works like 2.5D Visual-Sound (Gao and Grauman, 2019). More recent research has shifted towards direct, end-to-end synthesis from more abstract or multimodal inputs. Diffusion and flow-matching models are at the forefront of this trend, capable of generating high-fidelity FOA or binaural audio directly from text prompts, images, class labels, and explicit spatial positions (e.g., ImmerseDiffusion (Heydari et al., 2025), SonicMotion (Templin et al., 2025), OmniAudio (Liu et al., 2025a)). Transformer-based models excel at integrating complex, heterogeneous data streams; for instance, ViSAGe (Kim et al., 2025) generates FOA audio from video combined with camera position metadata, while ISDrama (Zhang et al., 2025) synthesizes expressive binaural speech from a rich mix of video, audio, text, and positional data. Other architectures serve specialized functions: VAEs are often used to learn disentangled latent representations for flexible spatial manipulation or to generate intermediate outputs like impulse responses (IRs) from 360° images (Kim et al., 2019), while GANs can incorporate detailed geometric data like 3D meshes to generate physically accurate binaural IRs, as demonstrated by Listen2Scene (Ratnarajah and Manocha, 2024).

D Discussion on Generation Architectures and Frameworks

The cascaded architecture is a postprocessing pipeline that upmixes a mono signal to add spatial dimensions. It is modular and controllable because content creation is separate from spatialization, letting researchers optimize rendering and use any mono input. Its quality is capped by the input, and the separation makes it hard to generate effects tightly tied to source physics, such as motion or deformation. While the end to end architecture synthesizes a full spatial sound field directly from high level inputs such as text, video, or 3D geometry, without relying on an existing audio signal. It offers strong creative potential and a more natural fusion of spatial attributes with content, enabling novel immersive soundscapes. The approach is more complex. Black box behavior hinders precise control of attributes such as timbre, and training usually needs large, well aligned multimodal datasets. The shift from cascaded to end to end marks a move from reprocessing to original creation.

Framework choice sets performance and scope. Early cascaded models for visual guided spatialization used CNNs such as U-Net(Ronneberger et al.,

Model	Input Format	Output Format	Framework
Xu et al. 2021	Mono; Image	Binaural	Diffusion-based
Binauralgrad (Leng et al., 2022)	Mono	Binaural	
Moûsai (Schneider et al., 2023)	Text	Binaural	
See-2-Sound (Dagli et al., 2024)	Image; (Text)	*Multi*	
Evans et al. 2024b	Text; (Audio; Duration)	Binaural	
DualSpec (Zhao et al., 2025)	Text	Binaural	
ImmerseDiffusion (Heydari et al., 2025)	Text; (Position)	FOA	
SonicMotion (Templin et al., 2025)	Text; Position	FOA	
Huang et al. 2022	Mono; Position	Binaural	Latent
Yang et al. 2022	Binaural/*Multi*	*Multi*	
Lee and Lee 2023	Mono; Position; Orientation	Binaural	Transformer-based
MusicGen (Copet et al., 2023)	Text	Mono/Binaural	
Ambisonizer (Zang et al., 2024)	Mono/Binaural	FOA	
ViSAGe (Kim et al., 2025)	Video; Camera Position	FOA	
ISDrama (Zhang et al., 2025)	Video; Audio; Text; Position	Binaural	
OmniAudio (Liu et al., 2025a)	360° Video	FOA	Flow Matching
Diff-SAGe (Kushwaha et al., 2025)	Class Label; Position	FOA	
Listen2Scene (Ratnarajah and Manocha, 2024)	3D Mesh; (Source & Listener Position)	Binaural IRs	GANs
SAGM (Li et al., 2024b)	Mono; Video	Binaural	

Table 5: Comparison of current spatial audio generative models. FOA denotes first-order ambisonics; *Multi* denotes multi-channel audio. Inputs/outputs in parentheses are optional. CNN-based models are omitted.

2015). CNNs are efficient for spectrogram like inputs but their local receptive fields limit long range temporal modeling. Transformers can handle long sequences and multimodal fusion. ISDrama(Zhang et al., 2025) integrates video, text, and position to generate long duration multi party dialogue, but quadratic complexity is costly so it adopts a Mamba Transformer. Diffusion and flow matching now lead on fidelity. ImmerseDiffusion(Heydari et al., 2025) and SonicMotion(Templin et al., 2025) show that iterative denoising yields highly realistic spatial soundscapes, at the cost of slow multi step sampling that is not ideal for real time. In short, CNNs suit lightweight feature processing, Transformers suit long sequence multimodal generation, and diffusion models suit non real time applications that aim for the highest perceptual quality.

E Dataset Details

Spatial audio data exists in a variety of formats, each reflecting different characteristics and tailored to specific tasks. Due to variations in recording equipment and application scenarios, spatial audio data comes in multiple formats, often accompanied by annotations and auxiliary data from other modalities. Moreover, because recording spatial audio is typically costly and resource-intensive, many existing approaches resort to using simulation systems to generate synthetic data from current monaural audio datasets. Some datasets also include real-world spatial audio crawled from the YouTube platform. The section focus on the acquisition and processing methods including both recorded and simulated data associated with various spatial audio formats, including multi-channel audio, First-Order Ambisonics, and binaural audio. A summary of commonly used datasets is presented in the Table 6.

F Objective Evaluation Metrics Details

F.1 Evaluation Metrics for Spatial Audio Understanding

SELD. The SELD task is evaluated using separate metrics for Sound Event Detection (SED) and Direction-of-Arrival (DOA) estimation. For SED, the one-second segment F-score and Error Rate (ER) are commonly used (Mesaros et al., 2016).

For DOA estimation, two frame-wise metrics are frequently employed (Adavanne et al., 2018b): **DOA Error** and **Frame Recall**. Let T be the total number of time frames. Denote by $\mathbf{DOA}_R^t$ the set of reference DOAs at frame t and by $\mathbf{DOA}_E^t$ the set of estimated DOAs. Define

$$D_R^t = \left|\mathbf{DOA}_R^t\right|, \quad D_E^t = \left|\mathbf{DOA}_E^t\right|.$$

Dataset	Format	Collect	Hours	Type	Labels
Sweet-Home (Vacher et al., 2014)	Multi	Recorded	47.3	Speech	Text
Voice-Home (Bertin et al., 2016)	Multi	Recorded	2.5	Speech	Text, Geomrtric
YT-ALL (Morgado et al., 2018)	FOA	Crawled	113	Audio	Video, Text
REC-STEEET (Morgado et al., 2018)	FOA	Recorded	3.5	Audio	Video
FAIR-Play (Gao and Grauman, 2019)	Binaural	Recorded	5.2	Audio	Video
SECL-UMons (Brousmiche et al., 2020)	Multi	Recorded	5	Audio	Text, Geometric
YT-360 (Morgado et al., 2020)	FOA	Crawled	246	Audio	Video
EasyCom (Donley et al., 2021)	Binaural	Recorded	5	Speech	Geometric, Text
Binaural(Richard et al., 2021)	Binaural	Recorded	2	Speech	Geometric
SimBinaural (Garg et al., 2023)	Binaural	Simulated	116	Audio	Video, Geometric
YouTube-Binaural (Garg et al., 2023)	Binaural	Crawled	27	Audio	Video
Spatial LibriSpeech (Sarabia et al., 2023)	FOA	Simulated	650	Speech	Text, Geometric
STARSS23 (Shimada et al., 2023)	FOA	Recorded	7.5	Audio	Video, Geometric
YT-Ambigen (Kim et al., 2025)	FOA	Crawled	142	Audio	Video
BEWO-1M (Sun et al., 2024)	Binaural	Simulated	2.8k	Audio	Text/Image, Geometric
MRSDrama (Zhang et al., 2025)	Binaural	Recorded	98	Speech	Text, Video, Geometric

Table 6: Comparison of current spatial audio datasets. FOA means first-order ambisonics, while Multi denotes multi-channel audio.

The **DOA Error** is defined as

$$\frac{1}{\sum_{t=1}^{T} D_E^t} \sum_{t=1}^{T} \mathrm{Hungarian}\big(\mathbf{DOA}_R^t, \mathbf{DOA}_E^t\big), \tag{3}$$

where $\mathrm{Hungarian}(\cdot,\cdot)$ denotes the optimal assignment cost computed by the Hungarian algorithm, using as the pairwise cost the central angle between a reference DOA (ϕ_R, λ_R) and an estimated DOA (ϕ_E, λ_E):

$$\sigma = \arccos\Big(\sin\lambda_E \sin\lambda_R + \cos\lambda_E \cos\lambda_R \cos|\phi_R - \phi_E|\Big). \tag{4}$$

Here $\phi \in [-\pi, \pi]$ is the azimuth and $\lambda \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ is the elevation.

To account for frames where the number of estimated DOAs does not match the number of reference DOAs, the **Frame Recall** is defined as

$$\text{Frame Recall} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\big(D_R^t = D_E^t\big), \tag{5}$$

where $\mathbf{1}(\cdot)$ is the indicator function, equal to 1 if its argument is true and 0 otherwise.

An ideal SELD method achieves an error rate of zero, an F-score of 1 (100%), a DOA Error of 0°, and a Frame Recall of 1 (100%). To compare submitted methods, each method is ranked individually for all four metrics, and final positions are determined by the cumulative minimum of these ranks.

The four cross-validation folds are treated as a single experiment: metrics are computed only after training and testing all folds. Intermediate measures (insertions, deletions, substitutions) are aggregated across folds before calculating the final metrics, rather than averaging per fold (Forman and Scholz, 2010).

Spatial Audio Separation. Metrics to measure the quality of separation, usually adopt the widely-used mir eval library metrics: Signal-to-Distortion Ratio (SDR) measures both interference and artifacts, Signal-to-Interference-Ratio (SIR) measures interference. Higher values indicate a better degree of separation (Ye et al., 2024).

Joint Learning. In joint learning, they typically employ two binary-classification-based evaluation metrics(Morgado et al., 2020). **AVC-Bin** (Audio–Visual Correspondence) determines whether an audio–video clip pair originates from the same video instance. **AVSA-Bin** (Audio–Visual Spatial Alignment) assesses the spatial consistency between the audio and visual streams.

For semantic segmentation, the model's dense-prediction capability is evaluated using **pixel accuracy and mean Intersection-over-Union (mean IoU)**. Additionally, **clip-level accuracy** is employed for action recognition.

F.2 Evaluation Metrics for Spatial Audio Generation

Monaural-to-Binaural Audio Generation. To comprehensively assess the fidelity of the synthesized binaural signal $\hat{x}$ concerning the reference binaural recording x, previous works (Leng et al., 2022; Liu et al., 2022) on **monaural-to-**

binaural audio generation adopt both objective and subjective criteria. Except for the perceptual measures, PESQ, all metrics are lower-is-better. Notation is unified as follows: $n \in \{1, \dots, T\}$ indicates time-domain samples; $c \in \{L, R\}$ indexes the two output channels; $k \in \{1, \dots, K\}$ and $m \in \{1, \dots, M\}$ denote STFT frequency and frame indices; $\mathrm{STFT}\{\cdot\}$ yields the complex time–frequency representation.

For **Wave $\mathbf{L_2}$**, The time-domain mean-squared error (MSE) captures sample-by-sample deviations:

$$\mathcal{L}_{L_2}^{\text{wave}} = \frac{1}{T} \sum_{n=1}^{T} \sum_{c \in \{L,R\}} \big(\hat{x}_c[n] - x_c[n]\big)^2. \quad (6)$$

Although it provides a well-behaved gradient and is easy to implement, it ignores the non-uniform frequency sensitivity of human hearing.

For **Amplitude $\mathbf{L_2}$**, after converting both signals to their magnitude spectra,

$$\begin{aligned} X(k,m) &= |\,\mathrm{STFT}\{x\}(k,m)|, \\ \hat{X}(k,m) &= |\,\mathrm{STFT}\{\hat{x}\}(k,m)|. \end{aligned} \quad (7)$$

The energy envelope mismatch is quantified as

$$\mathcal{L}_{L_2}^{\text{amp}} = \frac{1}{KM} \sum_{k=1}^{K} \sum_{m=1}^{M} \big(\hat{X}(k,m) - X(k,m)\big)^2. \quad (8)$$

For **Phase $\mathbf{L_2}$**, spatial cues rely strongly on interaural phase differences. To prevent phase-wrap artefacts, we minimize the wrapped phase distance:

$$\mathcal{L}_{L_2}^{\text{phase}} = \frac{1}{KM} \sum_{k=1}^{K} \sum_{m=1}^{M} \Big(\mathrm{wrap}\left(\angle \hat{X}(k,m) - \angle X(k,m)\right)\Big)^2, \quad (9)$$

where $\mathrm{wrap}(\theta) \in [-\pi, \pi)$.

To align perceptual quality with spectral accuracy, we average three complementary losses over a bank of M STFT configurations $\{\cdot^{(i)}\}_{i=1}^{M}$ as **Multi-Resolution STFT Loss (MRSTFT)**:

$$\begin{aligned} \mathcal{L}_{\text{SC}}^{(i)} &= \frac{\big\| \,|X^{(i)}| - |\hat{X}^{(i)}| \,\big\|_F}{\big\| \,|X^{(i)}| \,\big\|_F}, \\ \mathcal{L}_{\text{mag}}^{(i)} &= \frac{1}{N^{(i)}} \big\| \,|X^{(i)}| - |\hat{X}^{(i)}| \,\big\|_1, \\ \mathcal{L}_{\text{log}}^{(i)} &= \frac{1}{N^{(i)}} \big\| \log\big(|X^{(i)}| + \varepsilon\big) - \log\big(|\hat{X}^{(i)}| + \varepsilon\big) \big\|_1. \end{aligned} \quad (10)$$

$$\mathcal{L}_{\text{MRSTFT}} = \frac{1}{M} \sum_{i=1}^{M} \big(\mathcal{L}_{\text{SC}}^{(i)} + \lambda_{\text{mag}} \mathcal{L}_{\text{mag}}^{(i)} + \lambda_{\text{log}} \mathcal{L}_{\text{log}}^{(i)}\big). \quad (11)$$

This compound objective balances global spectral convergence with fine-grained magnitude fidelity across multiple time–frequency resolutions.

For **Perceptual Evaluation of Speech Quality (PESQ)**, the ITU-T P.862 standard maps symmetric (d_{sym}) and asymmetric (d_{asym}) perceptual distortions onto a MOS-like scale:

$$\text{PESQ} = 4.5 - 0.1\, d_{\text{sym}} - 0.0309\, d_{\text{asym}}, \quad (12)$$

yielding scores in $[-0.5, 4.5]$. Higher values denote closer perceptual similarity.

Wave/Amplitude/Phase L_2 losses provide gradient-friendly objectives that capture complementary signal aspects. MRSTFT augments them with multi-resolution spectral consistency. PESQ offers a single-ended perceptual estimate that correlates well with telecommunication speech quality. Together, this metric suite affords a balanced evaluation of both technical accuracy and perceptual realism in mono-to-stereo binaural conversion.

End-to-End Binaural Audio Generation. Evaluation metrics are highly varied for this task. In the case of binaural spatial audio, metrics can be computed based on interaural time difference (ITD), interaural level difference (ILD), and embeddings from a pretrained spatial-audio-understanding model(Zheng et al., 2024) to calculate specific performance indicators(Zhang et al., 2025). For the objective evaluation of IPD and ILD, they first convert the time-domain signal $x(n)$ into the frequency-domain signal $X(t, f)$ using the short-time Fourier transform (STFT):

$$\begin{aligned} X_i(t,f) &= \sum_{n=0}^{N-1} x_i(n) \cdot w(t-n) \cdot e^{-j2\pi f n}, \\ i &\in \{1, 2\}, \end{aligned} \quad (13)$$

where $w(t-n)$ is a window function, N is the window length, and i indicates the channel of the binaural audio. Next, they calculate the mel-spectrogram, IPD, and ILD based on the frequency-domain signals $X_i(t, f)$. The mel-spectrogram for each channel is calculated as:

$$S_i(t,m) = \log\left(|X_i(t,f)|^2 \times \text{melW}\right), \quad (14)$$

where melW is an M-bin mel filter bank. **IPD** is derived from the phase spectrograms of the left and right channels:

$$IPD(t,f) = \angle \frac{X_2(t,f)}{X_1(t,f)}. \tag{15}$$

Then, **ILD** is extracted from the loudness spectrum of the left and right channels:

$$ILD(t,f) = 20\log_{10}\left(\frac{|X_2(t,f)|+\varepsilon}{|X_1(t,f)|+\varepsilon}\right), \varepsilon = 1e^{-10}. \tag{16}$$

They calculate Mean Absolute Error (MAE) metrics based on the IPD and ILD extracted from the ground truth (GT) and the predicted speech. Since the IPD here is in radians and the ILD uses log10, the resulting values are quite small, especially after averaging the MAE over the time dimension. So, they multiply by 100 to make the results more intuitive.

Additionally, they analyze angular and distance metrics using SPATIAL-AST (Zheng et al., 2024). SPATIAL-AST encodes **angle and distance** embedding for binaural audio. They compute and average the cosine similarity for each 1-second segment based on the GT and predicted audio.

End-to-End FOA Generation. Current methods usually assess spatial localization accuracy by measuring azimuth error, elevation error, distance error, and spatial-angle difference (Heydari et al., 2025). Codec quality is evaluated via STFT and Mel distances between original and reconstructed FOA audio on the test set, using AuraLoss with default settings (Evans et al., 2024a,b). Plausibility of generated clips is quantified by the **Fréchet Audio Distance (FAD)** between generated and reference embeddings, and by **KL divergence** computed with a pretrained ELSA model. The **CLAP score**, the cosine similarity between spatial text embeddings and corresponding audio embeddings, is also reported. For the parametric model, KL divergence and CLAP are computed using spatial captions from the test set, despite training on non-spatial captions and parameters.

To measure spatial accuracy, they compare ground-truth and estimated **azimuth θ, elevation ϕ, and distance** d. Intensity vectors I_x, I_y, I_z are obtained by multiplying the omnidirectional channel W with the directional channels X, Y, Z:

$$I_x = W \cdot X, \quad I_y = W \cdot Y, \quad I_z = W \cdot Z \tag{17}$$

$$\theta = \tan^{-1}\frac{I_y}{I_x}, \quad \phi = \tan^{-1}\frac{I_z}{\sqrt{I_x^2 + I_y^2}}, \tag{18}$$

$$d = \sqrt{I_x^2 + I_y^2 + I_z^2} \tag{19}$$

They report the L1 norm of the differences for azimuth, elevation, and distance. For azimuth, they use the circular difference:

$$\mathrm{L1}_\theta = ||(|\theta - \hat{\theta}|, 2\pi - |\theta - \hat{\theta}|)||_1 \tag{20}$$

Spatial-angle error $\Delta_{\mathrm{Spatial-Angle}}$ is defined as (Van Brummelen, 2012):

$$a = \sin^2\left(\tfrac{\Delta_\phi}{2}\right) + \cos(\phi)\cos(\hat{\phi})\sin^2\left(\tfrac{\Delta_\theta}{2}\right) \tag{21}$$

$$\Delta_{\mathrm{Spatial-Angle}} = 2\arctan 2\left(\sqrt{a}, \sqrt{1-a}\right) \tag{22}$$

Here, Δ_ϕ and Δ_θ denote the linear and circular differences for elevation and azimuth, respectively.

G Subjective Evaluation Metrics Details

Mean Opinion Score (MOS) MOS delivers the gold-standard human judgment. It is obtained by averaging listener ratings over a five-point Likert scale:

$$\mathrm{MOS} = \frac{1}{N}\sum_{i=1}^{N} s_i, \tag{23}$$

where s_i is the score from the i-th participant. MOS serves as the definitive benchmark to which all objective metrics are ultimately calibrated.

MUSHRA MUSHRA (Series, 2014) targets systems with medium impairments. Listeners hear a visible reference, a hidden reference, one or more low quality anchors, and system outputs, then rate Basic Audio Quality on a 0–100 scale. The hidden reference checks reliability and anchors calibrate the scale. For spatial audio, dimensions can include spatial impression and stereophonic image quality. Trained experts are more sensitive to spatial artifacts, so MUSHRA is effective for assessing localization and immersion.

A/B and ABX A/B and ABX test (Boley and Lester, 2009) are also widely used. A/B asks listeners to compare two samples by a chosen criterion, such as realism or spatial impression. ABX asks whether an unknown sample X matches A or B to test perceptible differences. A/B is less fine grained than MUSHRA but is efficient for pairwise comparisons, validating improvements, and assessing specific perceptual dimensions.

H Further Discussion on Dataset and Evaluation Metrics

H.1 Training view: evolution and challenges from real recordings to large scale simulation

The evolution of spatial audio datasets shows the tension between the need for scale in deep learning and the high cost of real world recording. Early key datasets, such as EasyCom(Donley et al., 2021) for binaural separation and FAIR-Play(Gao and Grauman, 2019) for visual guided spatialization, provide valuable real recordings. They have high fidelity and include complex crosstalk and environmental noise in natural scenes. They are important for testing robustness. Their scale is small, often only a few to a few dozen hours, and scene diversity is limited. This is not enough for modern models that need hundreds of thousands of samples. To break this bottleneck, the community turned to large scale simulation. Spatial LibriSpeech(Sarabia et al., 2023) is a representative example. It convolves a mono corpus such as LibriSpeech with synthetic room impulse responses and produces more than 650 hours of training data with precise spatial labels. This approach offers unmatched scalability and precise control of acoustic parameters such as room size and reverberation time. It also introduces a core challenge, the sim to real gap. Models trained only on simulated data can drop in performance in real environments, since simulation cannot fully capture real acoustic propagation. To bridge this gap, real recording datasets such as RealMAN and RealImpact were created. They provide benchmarks to assess the realism of simulation and support a sim2real training paradigm, pretrain on simulation and finetune on real data. At the same time, web-crawled datasets such as YT-360(Morgado et al., 2020) and YT-AMBIGEN(Kim et al., 2025) offer another large scale source. They are in the wild and cover very diverse scenes. The main challenge is quality control. Audio quality varies and audio video spatial alignment is often not guaranteed. Effective cleaning, filtering, and labeling are therefore crucial.

H.2 Evaluation view: balancing objective metrics and subjective perception

The evaluation of spatial audio balances objective metrics and subjective tests. Objective metrics are repeatable and low cost. They are the mainstream for comparing models in research. For understanding tasks, the SELD metrics in the DCASE challenges(Mesaros et al., 2017) began with four separate metrics: error rate, F1, DOA error, and frame recall. On STARSS23(Shimada et al., 2023) they evolved to more integrated measures, such as location dependent F1 with angle and distance thresholds and relative distance error. This shows a move toward metrics that reflect overall task performance. For generation tasks, in addition to signal fidelity metrics such as SDR, researchers use spatial cue errors such as MAE of IPD and ILD, and proxy measures such as Fréchet Audio Distance and CLAP score to estimate perceptual quality and semantic consistency. Objective metrics have limits. They often cannot fully predict human listening experience. A system with a high SDR can still sound unnatural or not immersive. Subjective tests remain the gold standard for perceptual quality. The mean opinion score and the ITU MUSHRA test(Series, 2014) directly measure listener perception. MUSHRA uses a hidden reference and anchors and provides finer and more reliable scores than MOS for systems with medium impairments. Subjective tests are costly and hard to standardize. They are essential to calibrate objective metrics and to understand real perceptual strengths and weaknesses. An important future direction is to design new automatic objective metrics that align better with human perception. This can help close the gap between objective computation and subjective experience.

I Licenses and Availability

We respect the original licenses of all referenced artifacts and do not redistribute them. This survey does not create new deployable systems or redistribute data. Any consultation of third-party artifacts is limited to research/read-only use and complies with their intended-use statements and access conditions.