A Multimodal Evaluation Framework for Spatial Audio Playback Systems: From Localization to Listener Preference

Anonymous Author(s)
Submission Id: 4544

ABSTRACT

Spatial audio playback is a key feature of audio systems, shaping the core experience in immersive scenarios. However, objective evaluation methods for perceptual dimensions like sound field and sound image remain underdeveloped, hindered by the lack of fine-grained spatial audio datasets and the neglect of echoes and reverberation in diverse playback conditions. To address these challenges, we propose MESA, a multi-modal evaluation framework for spatial audio systems, and introduce PSA-MOS, a high-quality multi-scene spatial audio dataset. Specifically: 1) PSA-MOS provides 50 hours of high-quality spatial audio recordings spanning 6 playback scenarios and 7 device types, with detailed localization annotations and fine-grained MOS ratings across four perceptual dimensions. 2) We develop SAE-Encoder, a novel spatial audio encoder that captures both acoustic-spatial cues and fine-grained perceptual patterns. 3) MESA integrates visual scene context to enhance evaluation robustness through echo and reverberation modeling. Experimental results demonstrate that SAE-Encoder achieves superior performance in SELD tasks. With a two-stage training strategy, MESA exhibits strong correlation with human perceptual assessments on PSA-MOS, effectively guiding spatial audio quality optimization. The dataset and demos are available at http://mesa-demo.github.io.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; • **Applied computing → Arts and humanities**.

KEYWORDS

multi-modal evaluation, spatial audio playback systems, spatial audio dataset, perceptual-objective alignment

1 INTRODUCTION

With advances in spatial audio technology, applications ranging from VR &AR [4, 14] systems to automotive entertainment platforms [7] allow users to enjoy an immersive experience, which has created a critical need for comprehensive evaluation methodologies to assess the spatial audio playback capabilities of audio systems. However, subjective auditory testing methods are resource intensive and inefficient for the fast testing needs of audio playback systems. Thus, there is growing urgency to develop automated quality assessment frameworks for spatial audio that bridge objective metrics with human perceptions.

Substantial advancements have been achieved in the field of monaural audio quality assessment. As the gold standard for evaluating audio quality, human subjective evaluation has inspired a wide range of methodologies, ranging from traditional methods like PESQ [38], SI-SDR [23] to recent deep learning-based models like DNSMOS [36], and w2vMOS [11]. These methods focus on human perception and Quality of Experience (QoE) [34], aiming to align evaluations with subjective assessments [42].

Compared to the assessment of monaural audio quality , research on spatial audio[1] quality evaluation remains limited. Early works relied on binaural auditory cues, such as ITD(Interaural Time Differences) and ILD(Interaural Level Differences), to assess spatial quality [15, 20, 39]. As deep learning technology develops, researchers focus on the spatialization, providing assessment schemes based on localization estimation [27]. Later, evaluations expanded to include auditory effects like sound quality and distortion [28, 37]. Recently, Biswas et al. [5] and Eurich et al. [13] advanced efforts to unify the evaluation of monaural and binaural audio. However, these works fail to account for physical environment-related information, such as echoes and reverberation, in diverse audio playback environments. Additionally, there is still a lack of spatial audio corpora with fine-grained subjective ratings for evaluation models.

We attribute the challenges of spatial audio quality assessments to three aspects: 1) the lack of quantitative **evaluation datasets** based on auditory perception, 2) the ability to **integrate physical environment-related information** into evaluations, and 3) the establishment of **evaluation metrics** with fine granularity. And our work presents the first exploration to address these issues.

Spatial audio served for immersive scenarios are now commonplace. However, subjective evaluations of spatial audio playback quality are often limited to natural language descriptions. While some datasets provide subjective evaluations in a comparative quality manner [28], the significant human resources required for listening tests pose a major barrier to create quantified subjective evaluation datasets for replayed spatial audio. To advance spatial audio evaluation models, we developed PSA-MOS, the first spatial audio dataset for spatial audio playback systems. PSA-MOS comprises over 50 hours of recordings and more than 5,000 spatial audio samples, captured across six distinct spatial audio environments and played back using seven different types of audio devices. It also provides objective annotations such as sound events, in-door scenes and virtual sound source localization, along with fine-grained subjective MOS evaluations across four dimensions. To ensure accuracy and consistency, test sets from PSA-MOS are reviewed by experts, and subjective evaluation scores undergo consistency checks.

[1] For clarity, "spatial audio" refers to "stereo audio" with spatial context.

MM '25, October 27–31, 2025, Dublin, Ireland

ACM ISBN 978-1-XXXX-XXXX-X/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

Previous spatial audio evaluation models based on human perceptual assessment have achieved basic alignment with subjective evaluations [15, 28, 31]. However, they fail to incorporate audio playback scene information and lack fine-grained evaluation metrics, limiting their effectiveness in optimizing spatial audio playback in real-world scenarios. To address these challenges, we introduce MESA, a **M**ulti-modal **E**valuation Framework for **S**patial **A**udio playback systems. To align subjective and objective assessments, we designed a Transformer-based **S**patial **A**udio encoder for **E**valuation, named SAE-Encoder, as the acoustic front-end of MESA. This encoder not only performs audio understanding tasks such as sound event detection but also delivers fine-grained spatial audio evaluations. Additionally, MESA integrates vision-audio fusion to generate audio evaluations tailored to the acoustic environment, offering a comprehensive solution for incorporating scene-specific playback information. To better achieve the alignment between subjective and objective evaluations in our model, we also propose a two-stage training strategy for MESA.

In our experiments, we find that MESA maintains consistency between subjective and objective evaluations on fine-grained metrics and efficiently leverages scene information through vision-audio fusion. Using our two-stage training strategy, the SAE-Encoder alone demonstrates strong performance across objective understanding tasks like SELD. Furthermore, integrating the SAE-Encoder into the MESA framework results in high correlation with perceptual assessments on fine-grained dimensions such as sound quality, sound field, and sound image.

Overall, our key contributions are summarized as follows:

- We present PSA-MOS, a high-quality, multi-scene, and free-to-use spatial audio dataset collected from diverse spatial audio playback devices, accompanied by precise spatial annotations and fine-grained MOS evaluations to support the training of spatial audio assessment models.
- We propose SAE-Encoder, a novel spatial audio encoder architecture that not only extracts the content and spatial information of spatial audio but also effectively performs perceptual evaluation of spatial audio.
- We introduce MESA, a multi-modal evaluation framework for replayed spatial audio in automobiles. It incorporates the SAE-Encoder and leverages vision-audio fusion to achieve fine-grained evaluations of in-car spatial audio, effectively guiding improvements in spatial audio playback systems.

2 RELATED WORK

2.1 Spatial Audio Understanding

With the advancement of deep learning technologies, researchers have begun to explore the spatial audio field in depth using learnable approaches [30]. Due to the spatial characteristics of binaural audio, binaural audio localization [2] naturally became the first area of spatial audio to be systematically studied. Through the learning and exploration of ITD and ILD, researchers have been able to achieve sound source localization in both single-source [21] and multi-source[32] scenarios. Subsequent works [43, 44] adapted architectures like GRU [9] and Conformer [17] to achieve strong performance. To address the challenges of localization tasks caused by the lack of binaural audio data with precise localization, researchers have developed acoustic simulation techniques and algorithms that leverage spatial audio [8]. To study the impact of the surrounding physical environment on listener's auditory perception, researchers also model Room Impulse Response (RIR) functions to assist in spatial audio understanding tasks [24, 35]. Additionally, the evolution of large language models (LLMs) [18] has enabled their application beyond textual data, giving rise to multi-modal language models capable of performing spatial audio understanding tasks [41, 46].

2.2 Audio Quality Assessment

In monaural audio evaluation, studies primarily focus on QoE, employing non-intrusive assessment methods [19, 40] to approximate actual user experiences [45]. To tackle the dependency of traditional objective evaluation methods on corresponding clean references, researchers have explored using subjective MOS ratings [26] or designing evaluation frameworks [25, 29] that do not rely on paired audio references, thereby mitigating the need for high-quality audio pairs. As for spatial audio, the earliest works often learn metrics to evaluate spatial audio quality from known binaural auditory cues, such as ITD, ILD, and IACC (Inter-Aural Cross-Correlation) [15, 39]. Building on this foundation, DPLM [27] serves as a full-reference deep perceptual spatialization metric, offering a differentiable approach to evaluate stereo audio. Additionally, Delgado et al. [12] used Directional Loudness Maps to assess spatial audio. However, these works only measure spatial information, neglecting the equally important aspect of sound quality in auditory experience. To meet the demand for a more multidimensional evaluation of spatial audio, Bagousse et al. [22] assess spatial audio quality from the perspectives of timbre, space, and defaults, and the following works [31, 37] provides evaluation metrics that consider both spatialization quality (SQ) and listening quality (LQ). SAQAM [28] further considers the influence of Head-Related Transfer Functions (HRTFs) in spatial audio evaluation. Researchers have also attempted to combine the assessments of monaural and binaural audio quality [13]. Nevertheless, existing works have not provided fine-grained evaluations for spatial audio and have focused solely on the audio itself, overlooking the spatial effects created by echoes and reverberation from the playback environment of spatial audio playback systems.

3 PSA-MOS

In this section, we formally introduce PSA-MOS, the first spatial audio dataset designed for evaluating spatial audio playback systems. For better understanding and evaluating, PSA-MOS comprises 50 hours of noise-free binaural audio data, which is sampled at 48kHz and quantized at 16 bits. It consists of stereo audio recordings gathered from seven distinct playback devices across six diverse recording environments to ensure the generalizability of the model. Furthermore, we provide comprehensive, mutual annotations for each audio segment including sound events, virtual sound source localization and scene information. In addition, organize evaluators to conduct fine-grained MOS evaluations for the audio. The pipeline for constructing PSA-MOS is shown in Fig. 1. For more details, please refer to Appendix A.

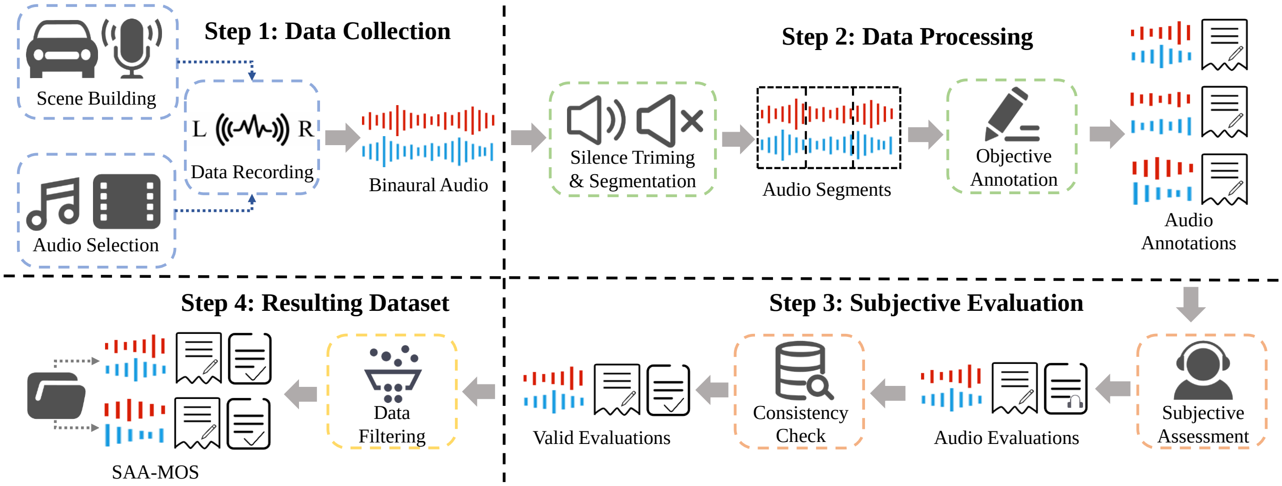


Figure 1: The pipeline for the construction of PSA-MOS.

3.1 Data Collection

In the selection of recording equipment and scenarios, we chose four representative automobile audio systems known for their spatial audio playback capabilities. Additionally, we purchased a spatial audio playback system that supports Dolby Digital 5.1-channel audio, as well as a binaural stereo system composed of two channels. Furthermore, during the recording process, we also removed the subwoofer from the Dolby Digital 5.1-channel audio system to use as another playback device. Next, we choose six distinct types of audio events — natural sounds, speech, music, song, traffic, and fighting sound — for testing and recording of the spatial audio systems. The entire recording process took place in both rented automobiles and custom-built private spatial audio environments without noise. The recorded audio data are saved in WAV format, sampled at 48 kHz, and quantized at 16 bits. For more details, please refer to Appendix A.1.

3.2 Data Processing

After recording the spatial audio, the data still falls short of high-quality standards in the following areas: 1) The raw audio contains numerous pauses and silences, which can interfere with spatial audio evaluation and increase computational overhead; 2) The audio lacks important annotations, such as audio events and virtual source directions, and individual clips are too long for efficient computation. To address the aforementioned issues, we designed a processing pipeline that includes **Silence Trimming, Segmentation, and Mutual Annotation** to improve the quality of the spatial audio data.

Silence trimming. After completing the data recording and initial division, lots of silent segments remain in the audio clips. To address this, we first conduct loudness normalization and then detect silent segments (dB < -40.0) longer than 200 ms and remove them. Cutting these silent segments allows for better evaluation of the playback capabilities of spatial audio systems and improves the accuracy and relevance of the annotations.

Segmentation. After silence trimming, we segment the audio into smaller fragments to facilitate training for evaluation tasks. Based on the time stamps of the silent segments, we define minimum and maximum duration intervals for the voiced regions and divide the audio clips accordingly. Each fragment is manually inspected to ensure it contains complete audio events and offers a full spatial audio experience. In the end, we obtained 9336 audio utterances.

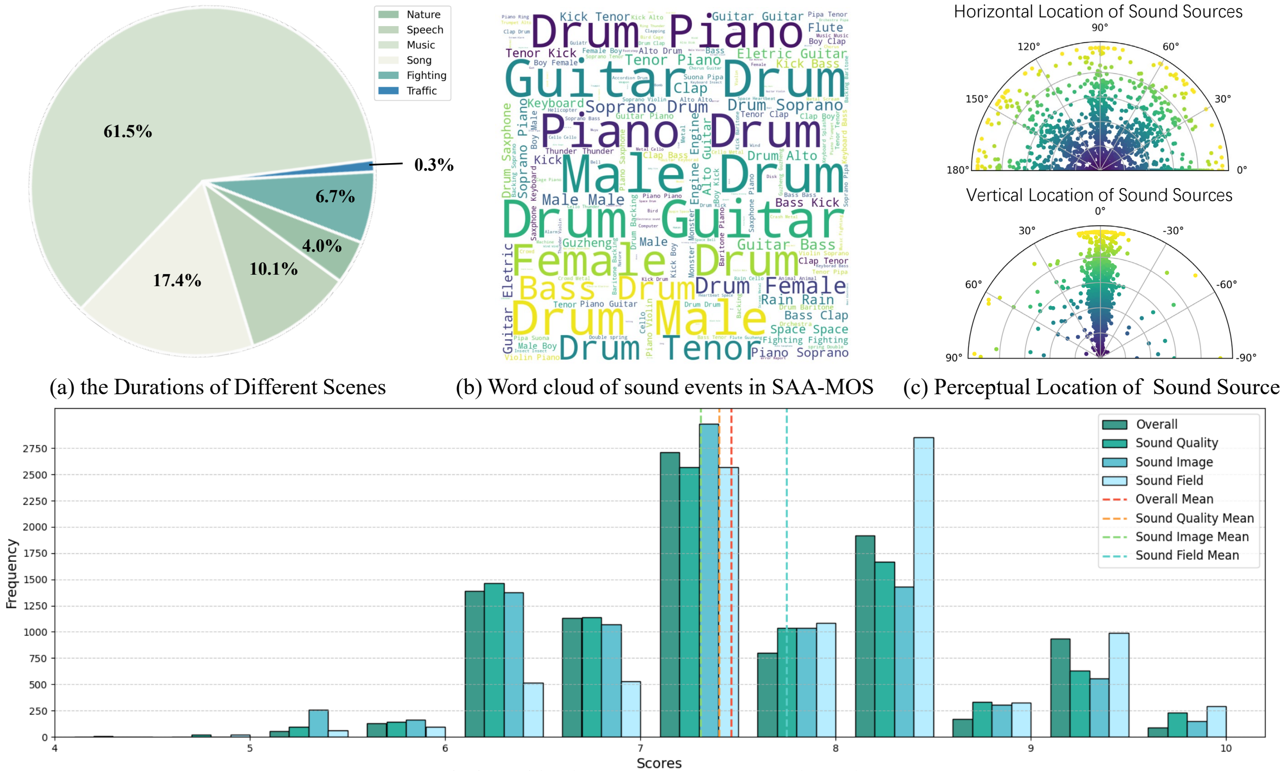
Annotation. We instruct our annotators to label information such as audio scenes, sound events, and the perceptual locations of vitrual sound source. The specific details of the annotations can be found in the Appendix A.2.

3.3 Subjective Evaluation

To evaluate the quality of the replayed spatial audio, we conduct a subjective evaluation with a team consisting of 20 audio experts. After the evaluation, we also check the validity and consistency of the scores. Details can be found in Appendix A.3.

Pipeline of Subjective Evaluation. Following the Mean Opinion Score (MOS) evaluation framework, we instruct listeners to assess the audio across four key dimensions: sound quality (SQ), sound field (SF), sound image (SI), and overall rating (OVRL). To minimize the influence of prior knowledge about the recording environment and equipment on subjective evaluations, and to further enhance the robustness and reproducibility of the results, the evaluation is conducted in a quiet indoor setting. Listeners use professional headphones and an audio interface, and are requested to rate various spatial audio samples on a Likert scale ranging from 1 to 10.

Validness and Concordance Test. TTo minimize interference from other factors and support repeated testing, we employ a recording and replay method during the evaluation phase, as mentioned above. However, this approach leads to an auditory experience that does not fully replicate the original playback environment. To ensure the validity of our evaluation method in relation to real-world listening experiences, we perform a Pearson correlation test [10] between the subjective scores of the spatial audio playback in the recording



(a) the Durations of Different Scenes (b) Word cloud of sound events in SAA-MOS (c) Perceptual Location of Sound Source

(d) the Subjective Scores across Four Dimensions

Figure 2: The statistical distribution of durations by scene, sound events, sound source locations and subjective evaluations.

environment and those of the replayed audio during the subjective evaluation process. Additionally, we perform a concordance test on the ratings provided by different listeners using the Kendall W coefficient of concordance [1] to ensure that all listeners adhere to the same rules and scales during the evaluation process.

3.4 Statistics

After the data collection and processing procedure, we check the audio quality and conduct a statistical evaluation, including the distribution of sound events, the distribution of virtual source directions and the distribution of ratings.

Distribution of Durations by Scene. As illustrated in Figure 2(a), PSA-MOS comprises six acoustic scenarios from two primary sources: music albums, and films. The Song and Music categories, which feature both high-frequency elements like piano and organ, as well as low-frequency components such as drum and bass, offer a comprehensive evaluation of spatial audio systems' playback capabilities. These form the core of PSA-MOS. To increase the diversity of our dataset, we also recorded natural soundscapes and daily life scenarios like speech, fighting, and traffic from movies.

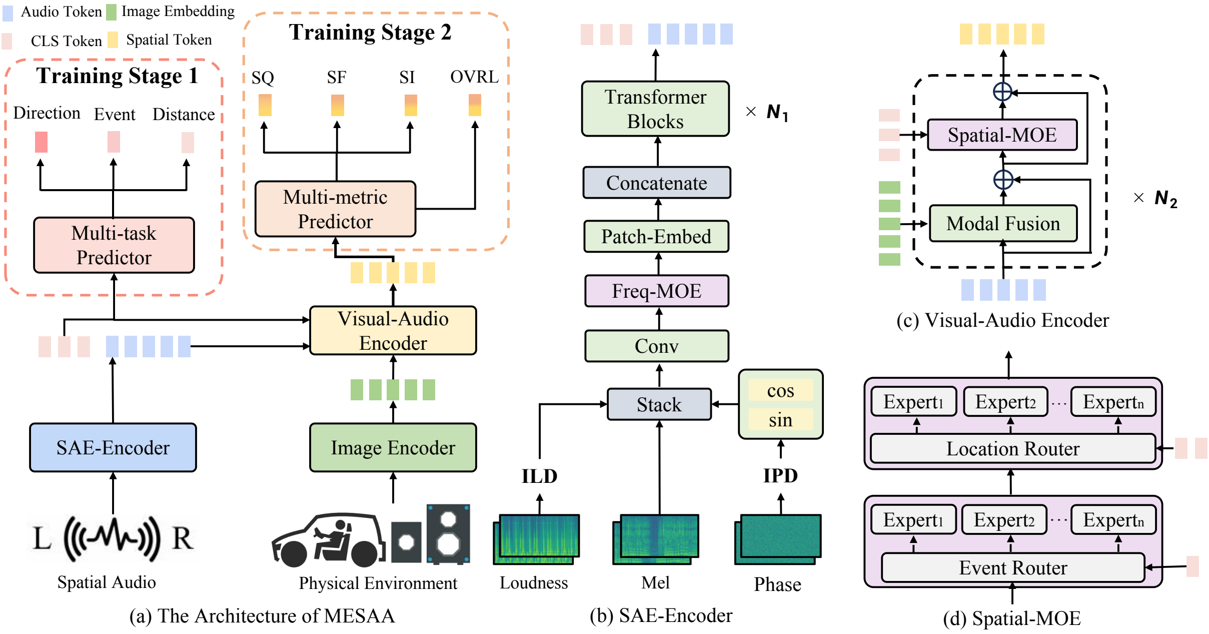
Distribution of Sound Events. Figure 2(b) demonstrates the variety of sound events in PSA-MOS. Since our recordings primarily focus on music, most sound events consist of instruments such as guitar, drums, and piano, along with various human voices. The dataset also includes nature sounds like rain and thunder, as well as mechanical sounds such as car engines and horns.

Distribution of Virtual Source Directions. Figure 2(c) shows the perceptual location of virtual sound source in the horizontal and vertical view. We can observe that the sound sources in PSA-MOS cover a wide range of directions and distances on the horizontal plane, providing a diverse auditory experience. However, in the vertical view, sound sources are concentrated on the horizontal plane, which may be due to the arrangement of the audio system.

Distribution of Ratings. Figure 2(d) demonstrates the distribution of the subjective scores in PSA-MOS. It shows that the ratings are primarily between 6 and 8, indicating that the spatial audio recordings we make are of high quality. In particular, the scores for sound field (SF) are relatively high, while the ratings for sound image (SI) are the lowest. This discrepancy may be attributed to the fact that current audio systems tend to prioritize delivering immersive sound effects over localization performance.

4 MESA

Fig. 3 illustrates the overall architecture of the MESA framework. We first introduce SAE-Encoder, our novel spatial audio encoder for evaluation, as the front-end of MESA. Then, we discuss the implementation of vision-audio fusion within MESA. Additionally, we will describe the overall framework of MESA, along with the two-stage training stragety. More details can be found in Appendix B.



(a) The Architecture of MESAA

(b) SAE-Encoder

(c) Visual-Audio Encoder

(d) Spatial-MOE

Figure 3: The overall architecture for MESA. In Figure(b), Freq-MOE represents Freqency-MOE and N_1 denotes the number of transformer blocks. In Figure(c), N_2 is the number of encoder layers.

4.1 SAE-Encoder

We propose a novel encoder based on transformer to capture spatial audio information including quality, spatiality and sound events. The model is shown on Fig. 3(b).

Feature Extractor. Our encoder extracts and integrates Mel Spectrogram, Interaural Phase Difference (IPD), and Interaural Level Difference (ILD) to capture features of spatial audio in terms of sound quality, sound field, and sound image. The Mel Spectrogram aligns with human hearing and captures audio information across multiple dimensions. IPD measures phase differences in the waveform, which helps model spatial info for low-frequency signals. ILD calculates the loudness difference between channels, which aids in modeling spatial info for high-frequency signals. As shown in Fig. 3(b), to tackle numerical instabilities, we apply cosine and sine transformations to the IPD to handle phase wraparound. Additionally, before calculating the ILD, we add a smoothing constant ε to ensure that the power spectrogram does not become zero. Both of IPD and ILD are then weighted by melW to align with Mel-spectrogram. The final front-end output, $\mathcal{Z}$, is a concatenation of these processed components as indicated in Equation 1. More details are in Appendix B.1.

$$\mathcal{Z} = [S_1; S_2; \text{ILD} \times \text{melW}; \cos(\text{IPD}) \times \text{melW}; \sin(\text{IPD}) \times \text{melW}] \tag{1}$$

Frequency-MOE. In spatial audio, high frequencies components, influenced by the outer ear and head, provide localization cues and enhance clarity and detail in auditory experience. Low frequencies content, while offering less directional information, are key for depth and immersion, shaping our sense of rhythm, distance and envelopment. To improve the encoder's feature extraction in frequency, we introduce Frequency-MOE (Mixture of Experts) in the SAE-encoder, selecting the best experts for different frequency ranges. Since the extracted features already include frequency information, we implement a block-based hard expert selection strategy in Frequency-MOE. For more details, please refer to Appendix B.1.

4.2 Visual-Audio Encoder

In the stage 2 training process, we integrate the visual modality and module into MESA to combine spatial environment information with audio. The model is shown on Figure 3(b).

Modal Fusion. Before fed into the visual-audio encoder, we first extract image features of panoramic physical environments through DINO [6] and obtain audio tokens using SAE-Encoder. Both the image features and audio tokens are then fed into a transformer block to get the hidden sequence. Specifically, we first pass through a modal fusion module that uses cross attention rather than a simplistic concatenation approach, allowing us to reason about how different image patches contribute to the audio after feature extraction. The equation is defined as follows:

$$\delta(V, \mathcal{Z}) = \text{Softmax}\left(\frac{\mathcal{Z}V^{\top}}{\sqrt{d_v}}\right)V \tag{2}$$

, where $\mathcal{Z}$ is the audio tokens, V is the visual features and d_v is the dimension of vision features. After that, we add the output from cross-attention to the audio tokens, apply a residual connection, and normalize the result using LayerNorm.

Spatial-MOE. Note that the auditory experience of spatial audio can vary significantly under different sound events and virtual sound source locations. To better capture the spatial information of binaural audio and leverage the echoes, reverberations, and other conditions provided by the panoramic view, we design Spatial-MOE in the visual-audio encoder to select the most suitable experts for handling different spatial audio scenarios, which is shown in Figure 3(d). Spatial-MOE consists of two expert groups: Event-MOE and Location-MOE, each comprising multiple experts. Event-MOE conditions on the first [CLS] token c_1, which represents the sound events in the audio, adjusting inputs to align with these events and selecting suitable experts, such as one specialized in open outdoor scenes. Location-MOE uses the other two [CLS] tokens, c_2 and c_3, which represent distance and direction, to select experts that are better suited for capturing spatial information, such as one specialized in extracting sounds like whispering near the ear. The routing strategy uses a dense-to-sparse Gumbel-Softmax [33], enabling dynamic and efficient expert selection. For more details, please refer to Appendix B.2.

4.3 Architecture

As illustrated in Figure 3, our model includes a spatial audio encoder, an image encoder, a visual-audio encoder, and two sets of multi-task predictors for spatial audio understanding and evaluation.

Image Encoder. Due to the strong spatial feature extraction capabilities of the DINO, which captures the complex structure of panoramic images through self-supervised learning and self-attention mechanisms, we directly utilize a pretrained DINO model based on ViT-B/16 [6], where the output dimension is 768 (d_p), aligning with the dimensions of audio tokens to facilitate modality fusion.

SAE-Encoder. The SAE-Encoder is a Transformer-based spatial audio evaluation model. It includes an audio extractor using Mel, IPD, and ILD, an MOE module to capture frequency information, and a Transformer encoder to extract quality and spatial cues.

Visual-Audio Encoder. The visual-audio encoder consists of relative position transformer blocks with MOE module. Specifically, it convolves a transformer encoder which includes multi-head cross-attention and two expert groups in MOE.

Predictors. Since our encoder has already extracted rich audio quality and spatial information, we use linear layers directly as the decoder for downstream tasks.

4.4 Two-Stage Training Strategy

To enable MESA to perform fine-grained evaluations of spatial audio while maintaining accurate objective perception, we designed a two-stage training strategy to promote the consistency between subjective and objective evaluations. In the first stage, the SAE-Encoder alone learns to understand the features of spatial audio through three tasks: sound event detection, distance prediction, and directional prediction. Then, we adopt the pre-trained SAE-Encoder and use PSA-MOS to enhance MESA's ability to perform fine-grained evaluation of replayed spatial audio. Details can be found in Appendix B.3.

Stage One Training. The SAE-Encoder is pre-trained to handle three tasks mentioned above. Following the approach in Spatial-AST [46], we apply cross-entropy loss for all three tasks. To discretize the prediction targets for distance and direction, we divide the distance into 0.5-meter intervals and the angles (both azimuth and elevation) into 1-degree intervals. The final loss function is as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{dis} + \lambda_3 \mathcal{L}_{doa} \tag{3}$$

, where $\mathcal{L}_{cls}$, $\mathcal{L}_{dis}$, $\mathcal{L}_{doa}$ represent the losses for detection, distance, and direction, while λ_1, λ_2, λ_3 are hyperparameters.

Table 1: This table presents a comparative analysis of various models using different features, focusing on key performance metrics: mean Average Precision (mAP, ↑), Error Rate at 20° (ER$_{20°}$, ↓), and Distance Error Rate (DER, ↓). Results for Spatial-AST are derived through direct inference using official checkpoints.

Baselines	Features		Metrics		
	IPD	ILD	mAP (↑)	ER$_{20°}$ (↓)	DER (↓)
SELDNet	✓	✗	42.71	25.19	38.40
Spatial-AST	✗	✗	50.86	30.94	28.60
	✓	✗	50.53	23.89	32.54
SAE-Encoder	✗	✗	**51.01**	29.73	28.91
	✓	✗	50.91	24.45	29.05
	✗	✓	50.65	27.72	28.35
	✓	✓	50.59	**24.12**	**28.29**

Stage Two Training. Since each audio clip is evaluated by only 20 experts, the final MOS scores have a minimum resolution of 0.05, which means that the MOS predictions are discrete. Using a regression model with L1 or L2 loss transforms the predicted values into a continuous distribution, which does not align with the habitual discrete scoring patterns of humans. What's more, it could be overly sensitive to outliers. On the other hand, directly using Cross-Entropy loss to construct a standard classification task fails to capture the ordinal nature of the scores, which contradicts the intrinsic logic of the rating task. To address these issues, we use Soft Ordinal Cross-Entropy, which incorporates ordinal information into the classification task. This approach explicitly models the order of scores and dynamically adjusts the penalties for errors based on their magnitude, as shown in Equation 4.

$$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} \left[- \sum_{i=0}^{C-1} \exp\left(- \frac{|y^{(n)} - i|}{s} \right) \cdot \log p_i^{(n)} \right] \tag{4}$$

, where N represents the number of samples in the batch, s is a scaling factor, and $p_i^{(n)}$ denotes the predicted probability of class i for the n-th sample. And the final loss function is provided below:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{OVRL} + \lambda_2 \mathcal{L}_{SQ} + \lambda_3 \mathcal{L}_{SI} + \lambda_4 \mathcal{L}_{SF} \tag{5}$$

, where λ_1, λ_2, λ_3 and λ_4 are hyperparameters.

5 EXPERIMENTS

In this section, we first introduce the experimental setup including dataset and model configurations. Then we report experimental results and conduct some analyses. Finally, we conduct ablation study for our loss function and the proposed Spatial-MOE. Fore more details, please refer to Appendix C.

Table 2: The experimental result of MESA. "Pre-train" indicates that we finetune the model following the SELD tasks. SQ represents Sound Quality, SI represents Sound Image, SF represents Sound Field, and OVRL represents overall score. For Wav2Vec 2.0, we design a naive Wav2Vec + Transformer architecture as the encoder, and all the prediction module in baselines are the same linear layers shown in Figure 3.

Baselines	Condition			OVRL		SQ		SF		SI	
	IPD	ILD	Pre-Train	L1(↓)	MP_{10}(↑)	L1(↓)	MP_{10}(↑)	L1(↓)	MP_{10}(↑)	L1(↓)	MP_{10}(↑)
Wav2Vec 2.0	-	-	✗	0.648	0.79	0.674	0.75	0.682	0.74	0.633	0.59
SAQAM	✗	✗	✗	0.297	0.86	0.288	0.86	0.299	0.84	0.366	0.66
SELDNet	✓	✗	✓	0.228	**0.90**	0.257	0.77	0.303	0.79	0.316	0.70
Spatial-AST	✓	✗	✗	0.211	0.84	0.242	0.80	0.279	0.87	0.279	0.66
	✗	✗	✓	0.216	0.86	0.261	0.80	0.278	0.81	0.279	0.71
	✓	✗	✓	0.209	0.86	0.242	0.86	0.280	0.89	0.276	0.69
MESA *w/o* vision	✓	✓	✗	0.221	0.80	0.253	0.77	0.280	0.82	0.287	0.65
	✓	✓	✓	0.198	0.89	**0.230**	**0.87**	0.272	0.82	0.274	0.65
MESA	✓	✓	✗	0.209	0.87	0.237	0.86	0.269	0.84	0.274	0.67
	✗	✗	✓	0.221	0.85	0.249	0.81	0.285	0.80	0.279	0.70
	✓	✗	✓	0.195	0.84	0.237	0.80	0.266	0.88	0.274	0.69
	✗	✓	✓	0.197	0.87	0.241	0.79	0.265	0.82	0.266	0.72
	✓	✓	✓	**0.194**	**0.90**	0.239	0.82	**0.254**	**0.91**	**0.261**	**0.78**

5.1 Experimental Setup

Dataset. In the first stage, to align with Spatial-AST [46], we use binaural audio simulated by SoundSpaces 2.0 [8], with sources filtered from AudioSet [16]. In the second stage, we train the evaluation model using PSA-MOS. All waveforms are processed into stereo audio which is sampled at 48kHz and quantized at 16 bits.

Baseline. For our encoder baseline comparsions, we select open-source models such as SELDnet [2] and Spatial-AST [46]. To ensure a fair comparison, SELDnet's feature extraction network is augmented with a 12-layer transformer block. For our evaluation comparisons, we choose SAQAM [28] as the baseline model, and we also build three evaluation models based on SELDnet, Spatial-AST, and wav2vec 2.0 [3] for compársions.

Metric. In the evaluation of SAE-Encoder, outlined in Table 1, we use mean Average Precision(mAP) for sound event detection; Error Rate@20° (ER_{20°) for Direction of Arrival (DoA) accuracy; and Distance Error Rate (DER) for measuring the accuracy of distance predictions. For evaluation of MESA, which shown in Table 2, we use L1 Loss in the test set to represent the accurancy of models. Besides, we use Precision@K to evaluate the model's performance, further measuring its perception ability. In the test set, we take randomly selected queries and calculate the number of correct class instances in the top K retrievals. And we finally report the mean Precision@10 (MP_{10}) over all test queries.

5.2 Performance of SAE-Encoder

Table 1 presents the experimental results for our proposed SAE-Encoder on SELD tasks, and we can observe the following: (1) Leveraging the Transformer encoder, both Spatial-AST and SAE-Encoder performs well in the sound event detection tasks, regardless of whether IPD and ILD are used. (2) With the inclusion of IPD, the model achieved a significant improvement in DoA accuracy, but its performance on distance predictions declined. And the inclusion of ILD enabled the model to gain better distance estimation ability. (3) With the usage of IPD and ILD, the model showed significant improvements in both DoA and distance predictions, achieving the lowest ER_{20° of 24.12 and DER of 28.29. This highlights the crucial role of loudness and phase information in spatial audio perception, especially for capturing spatial cues.

5.3 Performance of MESA

Table 2 presents MESA's experimental results in fine-grained evaluation predictions. We compare MESA against state-of-the-art models for evaluation tasks [3, 28] and SELD tasks [2, 46]. Additionally, we analyze the impact of audio features and visual inputs on evaluation performance. Finally, we discuss the effectiveness of the training strategy in achieving consistent evaluations.

Different Features' Effect. To support different audio features, we adjust the convolutional module's input channels to align with the dimensions of the final feature $\mathcal{Z}$. Incorporating IPD and ILD enhances spatial information, improving MESA's performance with an L1 loss of 0.254 for SI and 0.261 for SF. Notably, ILD, which captures variations from sound wave attenuation, has a greater impact on SI prediction as it relies on precise localization and high-frequency clarity. In contrast, IPD, which measures phase differences, is more sensitive to low-frequency spatial cues, allowing it to capture depth and envelopment, making it more critical for SF prediction. Furthemore, with the inclusion of spatial features, the MP_{10} scores also increases, which demonstrates that the model effectively captures and understands these spatial attributes. We also find that incorporating IPD and ILD improves the evaluation of SQ.

Table 3: The Ablation Study for loss function. We use MESA framework and all models are trained for 60 epochs.

Loss Function	OVRL		SQ		SF		SI	
	L1(↓)	MP_{10}(↑)	L1(↓)	MP_{10}(↑)	L1(↓)	MP_{10}(↑)	L1(↓)	MP_{10}(↑)
L1 Loss	0.328	0.80	0.357	0.77	0.303	0.79	0.317	0.55
Cross-Entropy	0.316	0.77	0.307	0.75	0.311	0.74	0.329	0.54
Soft Ordinal Cross-Entropy	**0.194**	**0.90**	**0.239**	**0.82**	**0.254**	**0.91**	**0.261**	**0.78**

Table 4: The Ablation Study for Spatial-MOE in MESA

	OVRL		SQ		SF		SI	
	L1(↓)	MP_{10}(↑)	L1(↓)	MP_{10}(↑)	L1(↓)	MP_{10}(↑)	L1(↓)	MP_{10}(↑)
MESA *ours*	0.194	0.90	0.239	0.82	0.254	0.91	0.261	0.78
w/o Sptial-MOE	0.209	0.84	0.263	0.77	0.287	0.76	0.292	0.73

Visual-Audio vs. Single Audio. For MESA *w/o* vision, we replace image embeddings with zero vectors, which prevents the model from learning spatial features during cross-attention. Experimental results indicate that integrating visual information enhances the model's capability to assess the spatial quality of audio, with significant improvements in both SF and SI. Specifically, the inclusion of the vision cues leads to a reduction in L1 Loss by 0.018 and 0.013 on the test set, accompanied by notable increases in MP_{10} scores, rising by 9% and 13%, respectively. However, in the Visual Audio Encoder, the primary focus is on spatial cues. This may cause the model to become overly reliant on spatial features, potentially overlooking acoustic details in the audio, leading to partial suppression of audio features. Consequently, in fine-grained evaluation predictions for SQ, the model's performance shows a slight decline.

Two-Stage Training vs. Direct Training. In the two-stage training approach, we fine-tune the spatial audio encoder on PSA-MOS, which has already developed spatial cue extraction capabilities through the SELD task. In contrast, direct training involves the model predicting scores on PSA-MOS without this prior fine-tuning. The results show that two-stage training significantly enhances the overall score prediction performance, particularly in spatial perception. Furthermore, in Top-k retrieval tasks, the two-stage training exhibits higher accuracy and stronger correlations, achieving MP_{10} scores of 0.91 for SF and 0.78 for SI. These findings indicate that our two-stage training strategy facilitates a deeper understanding of spatial cues within latent representations. Moreover, it is demonstrated that MESA successfully captures attributes and conducts evaluations in a human-like manner.

5.4 Ablation Study

Training Loss. To assess the effectiveness of Soft Ordinal Cross-Entropy, we performe a series of comparative experiments using various loss functions. The results, summarized in Table 5, lead to the following observations: 1) Standard classification models trained with Cross-Entropy loss exhibit a tendency to overfit, struggling to effectively extract evaluation-related attributes from the audio data. 2) Standard classification models using Cross-Entropy tend to overfit and fail to effectively extract evaluation attributes from the audio. 3) In contrast, Soft Ordinal Cross-Entropy achieves superior ordinal classification performance. It enables the model to predict accurate, fine-grained MOS evaluations while also demonstrating high accuracy in retrieval tasks.

Spatial-MOE. To validate the effectiveness of our Spatial-MOE module, we design a simple ablation experiment. For comparison, we replace the Spatial-MOE component in the Visual-Audio Encoder with a standard Feed Forward Network. As outlined in Table 4, the experimental results demonstrate that incorporating Spatial-MOE significantly enhances MESA's ability to evaluate spatial audio metrics, achieving improvements of 0.033 and 0.031 in the L1 Loss for SF and SI, respectively. Additionally, the inclusion of Spatial-MOE also positively impacts the overall evaluation of spatial audio and the perception of sound quality. These findings highlight the critical role of Spatial-MOE in refining the model's ability to capture nuanced spatial audio characteristics.

6 CONCLUSION

In this work, we presented MESA, a Multi-modal Evaluation Framework for Spatial Audio Playback Systems. To train and evaluate MESA, we introduced PSA-MOS, the first spatial audio dataset for audio systems featuring fine-grained subjective evaluations. We also proposed SAE-Encoder, a novel spatial audio encoder based on Transformer and MOE, which combines spatial audio perception with consistent subjective and objective evaluation capabilities. By leveraging cross-attention and the MOE structure, MESA utilizes spatial information from visual inputs of environments, achieving outstanding performance in spatial audio evaluation. PSA-MOS enables future research opportunities, such as virtual sound source localization and distance estimation, while MESA provides fine-grained and valuable guidance for improving spatial audio playback systems to create more realistic audio experiences.

REFERENCES

[1] Hervé Abdi. 2007. The Kendall rank correlation coefficient. *Encyclopedia of measurement and statistics* (2007).

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

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* (2020).

[4] Bradley J Baker. 2024. Virtual reality. In *Encyclopedia of Sport Management*. 1021–1023.

[5] Arijit Biswas and Guanxin Jiang. 2022. Stereo inse-net: Stereo audio quality predictor transfer learned from mono inse-net. *arXiv preprint arXiv:2209.11666* (2022).

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*.

[7] Jordan Cheer, Stephen J Elliott, and Marcos F Simón Gálvez. 2013. Design and implementation of a car cabin personal audio system. *Journal of the Audio Engineering Society* (2013).

[8] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip Robinson, and Kristen Grauman. 2022. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *Advances in Neural Information Processing Systems* (2022).

[9] Kyunghyun Cho. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[10] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing* (2009).

[11] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. 2022. Generalization ability of MOS prediction networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[12] Pablo M Delgado and Jürgen Herre. 2019. Objective assessment of spatial audio quality using directional loudness maps. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[13] BERNHARD EURICH, STEPHAN D EWERT, MATHIAS DIETZ, and THOMAS BIBERGER. 2024. A Computationally Efficient Model for Combined Assessment of Monaural and Binaural Audio Quality. *J. Audio Eng. Soc* (2024).

[14] Tira Nur Fitria. 2023. Augmented reality (AR) and virtual reality (VR) technology in education: Media of teaching and learning: A review. *International Journal of Computer and Information System (IJCIS)* 4, 1 (2023), 14–25.

[15] Jan-Hendrik Fleßner, Rainer Huber, and Stephan D Ewert. 2017. Assessment and prediction of binaural aspects of audio quality. *Journal of the Audio Engineering Society* (2017).

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

[17] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100* (2020).

[18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[19] Sandeep U Kadam, Vajid N Khan, Avaneesh Singh, Dattatray G Takale, Dattatray S Galhe, et al. 2022. Improve the performance of non-intrusive speech quality assessment using machine learning algorithms. *NeuroQuantology* (2022).

[20] Sven Kämpf, Judith Liebetrau, Sebastian Schneider, and Thomas Sporer. 2010. Standardization of PEAQ-MC: Extension of ITU-R BS. 1387-1 to multichannel audio. In *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*.

[21] 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* (2023).

[22] Sarah Le Bagousse, Mathieu Paquier, and Catherine Colomes. 2012. Assessment of spatial audio quality based on sound attributes. In *Acoustics 2012*.

[23] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. 2019. SDR–half-baked or well done?. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[24] 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* (2022).

[25] Pranay Manocha. 2024. *Do We Need a Reference Signal for Speech Quality Assessment?* Ph. D. Dissertation. Princeton University.

[26] Pranay Manocha and Anurag Kumar. 2022. Speech quality assessment through MOS using non-matching references. *arXiv preprint arXiv:2206.12285* (2022).

[27] Pranay Manocha, Anurag Kumar, Buye Xu, Anjali Menon, Israel D Gebru, Vamsi K Ithapu, and Paul Calamia. 2021. DPLM: A deep perceptual spatial-audio localization metric. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.

[28] Pranay Manocha, Anurag Kumar, Buye Xu, Anjali Menon, Israel D Gebru, Vamsi K Ithapu, and Paul Calamia. 2022. SAQAM: Spatial audio quality assessment metric. *arXiv preprint arXiv:2206.12297* (2022).

[29] Pranay Manocha, Buye Xu, and Anurag Kumar. 2021. NORESQA: A framework for speech quality assessment using non-matching references. *Advances in neural information processing systems* (2021).

[30] 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 (2010), 1–13.

[31] Miroslaw Narbutt, Jan Skoglund, Andrew Allen, Michael Chinen, Dan Barry, and Andrew Hines. 2020. Ambiqual: Towards a quality metric for headphone rendered compressed ambisonic spatial audio. *Applied Sciences* (2020).

[32] Thi Ngoc Tho Nguyen, Douglas L Jones, and Woon-Seng Gan. 2020. A sequence matching network for polyphonic sound event localization and detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 71–75.

[33] Xiaonan Nie, Xupeng Miao, Shijie Cao, Lingxiao Ma, Qibin Liu, Jilong Xue, Youshan Miao, Yi Liu, Zhi Yang, and Bin Cui. 2021. Evomoe: An evolutional mixture-of-experts training framework via dense-to-sparse gate. *arXiv preprint arXiv:2112.14397* (2021).

[34] Alessandro Ragano, Emmanouil Benetos, and Andrew Hines. 2023. Audio quality assessment of vinyl music collections using self-supervised learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[35] 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*.

[36] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2021. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

[37] Hualin Ren, Christian Ritz, Jiahong Zhao, and Daeyoung Jang. 2024. Towards an Objective Quality Metric for Interpolated Directional Room Impulse Responses. *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2024).

[38] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*.

[39] Jeong-Hun Seo, Sang Bae Chon, Keong-Mo Sung, and Inyong Choi. 2013. Perceptual objective quality evaluation method for high quality multichannel audio codecs. *Journal of the Audio Engineering Society* (2013).

[40] Kailai Shen, Diqun Yan, Jing Hu, and Zhe Ye. 2024. Non-intrusive speech quality assessment: A survey. *Neurocomputing* (2024).

[41] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Jun Zhang, Lu Lu, Zejun Ma, Yuxuan Wang, et al. 2024. Can Large Language Models Understand Spatial Audio? *arXiv preprint arXiv:2406.07914* (2024).

[42] Matteo Torcoli, Thorsten Kastner, and Jürgen Herre. 2021. Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021).

[43] Qing Wang, Li Chai, Huaxin Wu, Zhaoxu Nian, Shutong Niu, Siyuan Zheng, Yuyang Wang, Lei Sun, Yi Fang, Jia Pan, et al. 2022. The nerc-slip system for sound event localization and detection of dcase2022 challenge. *DCASE2022 Challenge, Tech. Rep.* (2022).

[44] Qing Wang, Huaxin Wu, Zijun Jing, Feng Ma, Yi Fang, Yuxuan Wang, Tairan Chen, Jia Pan, Jun Du, and Chin-Hui Lee. 2020. The USTC-iFlytek system for sound event localization and detection of DCASE2020 challenge. *IEEE AASP Chall. Detect. Classif. Acoust. Scenes Events* (2020).

[45] Xinwen Yue, Yupei Zhang, Jianqian Zhang, Zhiyu Li, Jing Wang, and Shenghui Zhao. 2024. Non-Intrusive Audio Quality Assessment Based on Deep Neural Network for Subjective MOS Prediction. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*.

[46] 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* (2024).