CSAVocoder: A Causal Spatial Audio Vocoder Towards Real-Time Spatial Audio Generation

Anonymous ACL submission

Abstract

Spatial audio vocoders aim to convert mel-spectrograms produced by generative models into spatial audio waveforms. Most existing vocoder research focuses on monaural audio, and direct extensions to spatial audio often degrade spatial quality by ignoring inter-channel cues. We present CSAVocoder, a causal GAN-based Spatial Audio Vocoder that jointly optimizes waveform fidelity and spatial rendering. Our framework introduces a spatial adaptor that fuses multi-channel mel-spectrograms with dynamic source-listener pose information, and a spatial consistency discriminator that explicitly supervises inter-channel spatial cues such as interaural level and phase differences. To meet real-time requirements, we design a strictly causal, stateful generator that supports efficient streaming inference with constant memory overhead. Experiments on large-scale spatial audio datasets demonstrate that CSAVocoder ensures audio quality and spatial fidelity while maintaining real-time performance. Our demo page is at: `https://csavocoder.github.io`.

1 Introduction

Unlike monaural audio, spatial audio renders sound sources at different directions and distances, providing a more immersive listening experience. It reconstructs a three-dimensional sound field and exploits the natural localization mechanisms of the human auditory system. By accurately modeling these cues, spatial audio delivers a strong sense of presence and realism in digital environments.

Spatial audio is increasingly important in applications such as virtual reality, augmented reality (Gupta et al., 2022; Kailas and Tiwari, 2021), and immersive gaming (Raghuvanshi and Snyder, 2018; Broderick et al., 2018; Yadegari et al., 2022). Recent generative models have made progress in spatial audio synthesis (Zhu et al., 2025; Lu et al., 2025), but many of them operate in the mel-spectrogram domain and rely on a vocoder to produce waveforms. Works such as ISDrama (Zhang et al., 2025a) and DualSpec (Zhao et al., 2025a) use pretrained HiFi-GAN-style vocoders and achieve high single-channel quality, yet they largely ignore inter-channel spatial consistency. Most vocoder studies still target single-channel audio, and direct extensions to spatial audio often degrade spatial quality because they ignore inter-channel cues so the necessity of relative pose between the sound source and the listener is a critical spatial factor in spatial audio rendering. Recent works (Heydari et al., 2025; Singh Kushwaha et al., 2024; Templin et al., 2025) use various forms of spatial information, including explicit coordinates and features extracted from visual inputs. The relative position controls loudness and spectral coloration, while orientation affects perceived direction and spatial awareness. Therefore, an effective spatial audio vocoder must explicitly model and exploit pose information to improve both signal quality and spatial perception.

On the other hand, real-time and efficiency requirements further complicate spatial audio rendering. In virtual and augmented reality, user interaction and rapid scene changes require spatial audio to react with low latency in order to maintain immersion. Prior work (Joy et al., 2024; Zhang et al., 2025a) emphasizes real-time rendering and the real-time factor (RTF). Since the vocoder is the final stage of spatial audio generation, its inference speed directly impacts end-to-end system latency and is crucial for real-time applications.

Designing a spatial audio vocoder that is both powerful and efficient is therefore challenging. The model must simultaneously (1) synthesize waveforms with high fidelity, (2) render perceptually valid spatial cues such as interaural level differences (ILD) and interaural phase differences (IPD), and (3) learn the complex mapping from pose to acoustic behavior, including source position and motion. In addition, the vocoder needs to be causal and support low-latency streaming inference that generates audio continuously in chunks.

To address these challenges, we propose CSAVocoder. In summary, our contributions are:

- We design a GAN-based spatial audio vocoder with a causal architecture that supports low-latency streaming inference while maintaining high-quality spatial audio synthesis.
- We introduce a pose-conditioning mechanism using position adaptor that encodes the relative source–listener pose and mel adaptor to capture inter-channel relationships, improving spatial audio rendering and perceptual quality.
- We propose an architecture that supports multiple spatial audio formats and learns an end-to-end mapping from multi-channel mel-spectrograms to multi-channel spatial audio waveforms.

2 Related Work

Our work lies at the intersection of spatial audio rendering, high-fidelity neural vocoders, and real-time synthesis.

2.1 Spatial Audio Rendering

Spatial audio rendering aims to construct immersive auditory scenes by modeling sound propagation in three-dimensional space. Among existing representations, binaural audio and First-Order Ambisonics (FOA) are particularly central. Binaural audio directly models ear-canal signals via head-related transfer functions (HRTFs) and is the final perceptual format for headphone playback, while FOA provides a spherical-harmonic, scene-centric representation with rotational equivariance and is widely used in VR and 360° video systems. These two formats are therefore the primary targets of many generative spatial audio models.

A broad line of work studies spatial audio generation from visual, textual, or multimodal inputs. 2.5D Visual Sound (Gao and Grauman, 2019) upmixes monophonic audio to binaural signals using visual cues in a regression setting. More recent methods move toward end-to-end spatial generation: ViSAGe (Kim et al., 2025) predicts FOA from silent video, ISDrama (Zhang et al., 2025a) models long-form spatial narratives with explicit real-time constraints, Diff-SAGe (Singh Kushwaha et al., 2024) applies diffusion in the complex spectral domain to better preserve inter-channel phase, and BEWO (Sun et al., 2024) enables text-driven binaural generation. ImmerseDiffusion (Heydari et al., 2025) and In-the-Wild Audio Spatialization (Pan et al., 2025) use spatial and semantic conditions to synthesize FOA or binaural audio for complex scenes.

Many of these systems operate primarily in the spectral domain and rely on separate vocoders or reconstruction stages, which introduce additional latency. Spatial information is often injected implicitly via latent variables or high-level prompts, and only a few works, such as ISDrama and ImmerseDiffusion, combine explicit spatial conditioning with considerations of real-time performance. This places strong requirements on the spatial audio vocoder at the end of the pipeline to generate high quality spatial audio with precise spatial cues.

2.2 Neural Vocoders

Neural vocoders map acoustic features to waveforms and form the last stage of audio generation. GAN-based vocoders dominate due to favorable quality-efficiency trade-offs. HiFi-GAN (Kong et al., 2020) introduces multi-period and multi-scale discriminators; BigVGAN (Lee et al., 2022) improves robustness via periodic activations and anti-aliasing; FARGAN (Valin et al., 2024), CARGAN (Morrison et al., 2022), and QGAN (Chaudhary and Abrol, 2024) reduce parameters and computing complexity. MusicHifi (Zhu et al., 2024) is an efficient high-fidelity stereophonic vocoder which can be used to enhance the fidelity of a low-resolution audio.

Alternative approaches operate in structured domains. Vocos (Siuzdak, 2024) predicts complex STFT coefficients; AF-Vocoder (Chen et al., 2025) applies frequency-domain artifact filtering; DisCoder (Lanzendörfer et al., 2025) generates in the latent space of neural audio codecs. Diffusion and flow-based vocoders such as DiffWave (Kong et al., 2021), Fregrad (Nguyen et al., 2024), and WaveFM (Luo et al., 2025) offer high perceptual quality via iterative denoising or direct transport learning. These existing vocoders primarily target monophonic or stereophonic audio and do not explicitly model spatial cues, limiting their effectiveness for spatial audio rendering.

2.3 Real-time Speech Synthesis

Real-time synthesis is critical for interactive applications where latency must stay below perceptual thresholds, favoring causal architectures and streaming inference. Online voice conversion systems such as CONAN (Zhang et al., 2025b) use chunk-wise state caching for bounded-delay conversion. For vocoders, WaveHax (Yoneyama et al., 2025b) and MS-WaveHax (Yoneyama et al., 2025a) adopt causal convolutions with shuffle-based upsampling; DLL-APNet (Du et al., 2025) combines distillation and simplification; MelFlow (Welker et al., 2025) adapts flow models to causal mel-to-waveform mapping; BinauralFlow (Liang et al., 2025) demonstrates streamable binaural generation. These advances motivate spatial vocoders that jointly achieve high spatial fidelity and streaming capability.

3 Method

3.1 Task Definition

We aim to synthesize a multi-channel spatial audio waveform $\mathbf{y} \in \mathbb{R}^{C\times L}$ from a multi-channel mel-spectrogram $\mathbf{M} \in \mathbb{R}^{C\times F\times T}$ and the corresponding spatial pose sequence $\mathbf{P} \in \mathbb{R}^{D_p\times T_p}$. Here, C denotes the number of channels, L is the waveform length, F is the number of mel frequency bins, and T is the number of mel frames. The sequence $\mathbf{P}$ captures the time-varying pose of the sound source relative to the listener, where D_p is the pose dimension and T_p is the number of pose samples. Each pose vector consists of a 3D Cartesian position (x, y, z) and a 4D quaternion (q_w, q_x, q_y, q_z) that encodes orientation, so $D_p = 7$.

We formulate the problem as learning a conditional generative function G that maps the inputs to the target waveform:

$$\mathbf{y} = G(\mathbf{M}, \mathbf{P}; \theta), \qquad (1)$$

where θ denotes the learnable parameters of the generator.

3.2 GAN-based Vocoder

Our framework is built on HiFi-GAN vocoder consisting of a generator G and a set of discriminators D, and extend its generator and discriminator stack to support spatial conditioning and strictly causal, streaming synthesis.

3.2.1 Generator

The generator follows the overall topology of HiFi-GAN which uses a convolutional network to upsample the input mel-spectrogram on temporal domain.

The Generator takes output from the Spatial Mel Adaptor and Spatial Position Adaptor as conditioning inputs. Tensors are fed into a series of upsampling and residual blocks to gradually increase the temporal resolution to that of the target waveform. We replace standard transposed convolutions with our ShuffleUpsampleBlock. First, the CausalConv1d block projects the channels from C to $C_{\text{out}} \cdot s$, producing $\mathbf{X}' \in \mathbb{R}^{B\times(C_{\text{out}}s)\times T_{\text{in}}}$. Then a ShuffleBlock reshapes this tensor to $\mathbf{X}'' \in \mathbb{R}^{B\times C_{\text{out}}\times(T_{\text{in}}s)}$ by folding extra channels into the time dimension. Since pixel shuffle is a pure tensor reordering without temporal mixing, it preserves the causality of the preceding convolution and yields artifact-free causal upsampling.

The residual blocks forming the multi-receptive-field fusion (MRF) stack are modified in the same spirit. Each StreamingResBlock consists of several causal convolutions with different dilation rates to capture patterns at multiple temporal scales, and maintains an internal buffer whose length matches its effective left context.

3.2.2 Discriminator

Conventional Wave and Spectral Discriminators To ensure high fidelity in both waveform and spectral domains, we adopt the standard MPD and MSD from HiFi-GAN (Kong et al., 2020) and MRD from BigVGAN (Lee et al., 2022) to ensure high-fidelity waveform and spectral reconstruction. Each sub-discriminator computes an STFT with a specific configuration, allowing the model to detect artifacts that appear only at particular time-frequency resolutions.

Spatial Consistency Discriminator To explicitly supervise spatial structure, we introduce a Spatial Consistency Discriminator (SCD) that operates on multi-channel log-mel spectrograms and provides spatially informed adversarial gradients to the generator. Given a multi-channel waveform $\mathbf{y} \in \mathbb{R}^{B\times C\times T}$, the SCD computes $\mathbf{M} \in \mathbb{R}^{B\times C\times F\times T'}$ and projects it via a 2D convolution into latent features $\mathbf{X} \in \mathbb{R}^{B\times d\times C\times T'}$. An axial-attention backbone then applies MHSA along the temporal axis $(B \cdot C, T', d)$ and along the channel axis $(B \cdot T', C, d)$, jointly modeling long-range dynamics and inter-channel relationships such as ILD/IPD in binaural signals and coherent patterns in FOA. A lightweight convolutional head finally maps the attended features to a scalar spatial consistency score per segment, complementing conventional discriminators that primarily target single-channel fidelity.

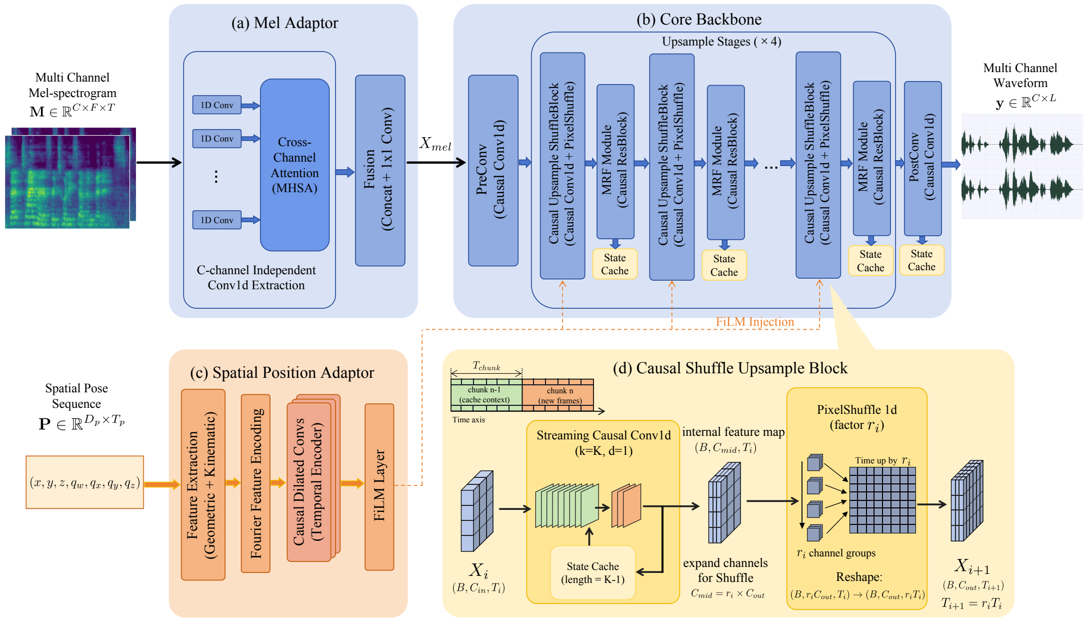

Figure 1: Overview of our model architecture.

3.2.3 Training Objectives

To train the generator G (the vector-field network v_θ) and the discriminator set D, we use a composite objective composed of several weighted loss terms. We adopt the standard Least-Squares GAN adversarial loss ($\mathcal{L}_{\text{adv}}$) (Mao et al., 2017), Feature Matching loss($\mathcal{L}_{\text{fm}}$) (Kumar et al., 2019), and multi-resolution spectral reconstruction losses ($\mathcal{L}_{\text{mel}}$ and $\mathcal{L}_{\text{STFT}}$) (Kong et al., 2020). The detailed formulations of these standard objectives are provided in Appendix B. Our primary contribution to the objective function is the format-aware Spatial Loss, designed to explicitly supervise spatial cues.

Spatial Loss Standard spectral losses treat channels independently, failing to constrain inter-channel spatial cues. We propose a format-aware spatial loss $\mathcal{L}_{\text{spatial}}$ that explicitly supervises physical attributes.

For Binaural Audio, based on the Duplex Theory, we combine Interaural Phase Difference (IPD) and Level Difference (ILD) losses: $\mathcal{L}^{\text{Bin}}_{\text{spatial}} = \lambda_{\text{IPD}}\mathcal{L}_{\text{IPD}} + \lambda_{\text{ILD}}\mathcal{L}_{\text{ILD}}$. Specifically, $\mathcal{L}_{\text{IPD}}$ operates on multi-resolution STFTs and compares phase differences in a sine-cosine embedding to avoid wrapping, with supervision concentrated in the low-frequency region using a Gaussian weighting. Conversely, $\mathcal{L}_{\text{ILD}}$ measures the discrepancy between log-magnitude level differences of the two ears, emphasizing high frequencies through a complementary weighting.

For FOA Audio, we define physical descriptors: $\mathcal{L}^{\text{FOA}}_{\text{spatial}} = \lambda_{\text{iv}}\mathcal{L}_{\text{iv\_dir}} + \lambda_{\text{r}}\mathcal{L}_{\text{r}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{elog}}\mathcal{L}_{\text{elog}}$. Direction-related terms ($\mathcal{L}_{\text{iv\_dir}}, \mathcal{L}_{\text{r}}$) constrain the intensity vector's angle and magnitude with a low-frequency bias, while diffusion-related terms ($\mathcal{L}_{\text{diff}}, \mathcal{L}_{\text{elog}}$) capture ambient envelopment with a mid-high-frequency bias.

To stabilize training, all terms are modulated by an energy-based soft mask derived from the ground-truth signal. Detailed formulations are in Appendix B.4.

Full Objective The total loss functions for the generator and the discriminators are defined as weighted sums of the components described above.

For each discriminator D_k in the discriminator set D, the total loss consists only of the adversarial term:

$$\mathcal{L}_D = \sum_k \mathcal{L}_{\text{adv}}(D_k; G). \tag{2}$$

For the generator G, the total loss is defined as

$$\begin{aligned}\mathcal{L}_G = {} & \mathcal{L}_{\text{adv}}(G; D) + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}} + \lambda_{\text{mel}}\mathcal{L}_{\text{mel}} \\ & + \lambda_{\text{STFT}}\mathcal{L}_{\text{STFT}} + \lambda_{\text{spatial}}\mathcal{L}_{\text{spatial}},\end{aligned} \tag{3}$$

where λ_{fm}, λ_{mel}, λ_{STFT}, and λ_{spatial} are hyperparameters that balance the contributions of different loss terms.

3.3 Causal Architecture for Streaming Synthesis

We redesign the HiFi-GAN generator as a fully causal, explicitly stateful architecture tailored for

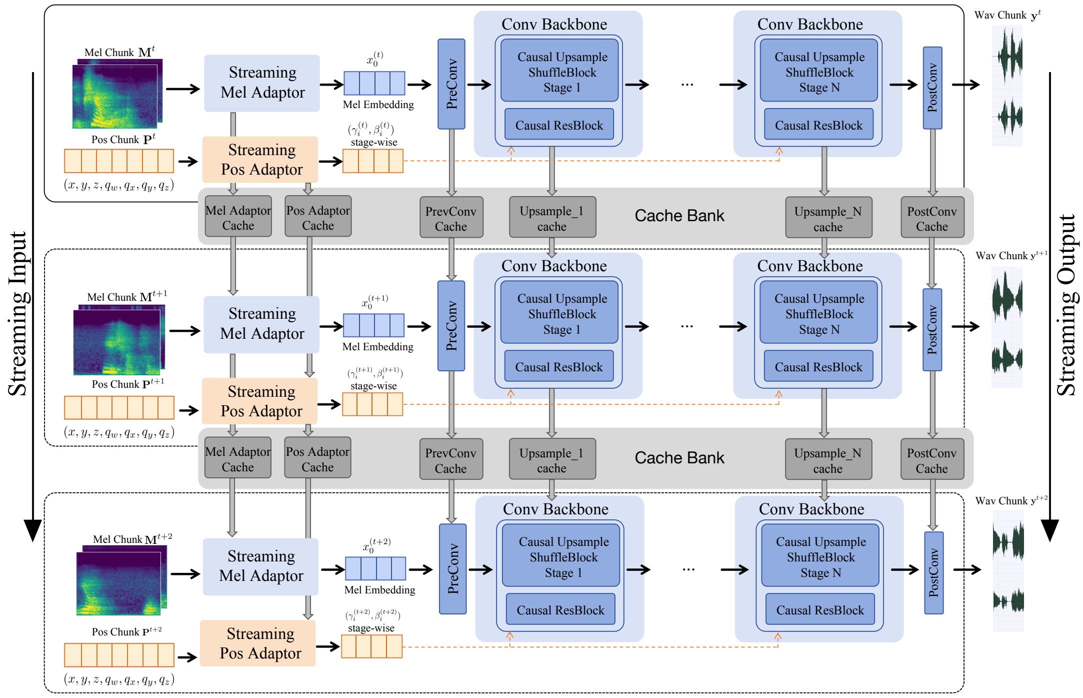

Figure 2: This figure shows the continuous streaming infer pipeline. Starting with multi-channel mel-spectrogram, we compute Mel Embedding and Pos information with Mel Adaptor and Position Adaptor. Then they are fed into the Conv Backbone and after upsampling and resblock, a chunk of wave is generated. All streaming blocks has its own state cache in the cache bank which restores a few chunk states before themselves and computes results strictly following causal restrictions.

streaming synthesis. All stages from mel features to waveform are constructed to satisfy strict causality, while a stateful inference mechanism avoids redundant computation in chunk-based processing.

Strict Causal Property When mapping a mel-spectrogram $M = \{m_1, \dots, m_T\}$ to waveform $W = \{w_1, \dots, w_{T'}\}$. , strict causality requires that each output sample w_t depends only on input frames $\{m_1, \dots, m_i\}$ whose timestamps do not exceed that of w_t. Any dependency on future frames m_j with $j > i$ violates this constraint. Our design enforces this property at the operator level.

Stateful Streaming Inference. Causality alone is insufficient for efficient streaming, since naively concatenating long contextual prefixes for each chunk leads to substantial redundant computation. We therefore implement all context-dependent layers in a stateful form, where each layer accepts both the current input chunk and a compact cache from the previous step, and returns the current output together with an updated cache that stores exactly the left-context features needed for the next chunk. As shown in Figure 2, during streaming synthesis, the generator processes a sequence of mel chunks while propagating a global state object that aggregates the caches of all stateful layers, avoiding any recomputation of past activations.

3.4 Spatial Adaptor

Standard mono-channel vocoders lack mechanisms to process multi-channel spectrograms or incorporate heterogeneous pose conditioning. To bridge this gap, we introduce the Spatial Adaptor, comprising two parallel modules to encode spectral and geometric cues respectively.

3.4.1 Attentional Mel Adaptor

This module fuses the multi-channel mel-spectrogram $\mathbf{M} \in \mathbb{R}^{B\times C\times F\times T}$ into a unified single-stream representation $\mathbf{X}_{\text{mel}} \in \mathbb{R}^{B\times d_{\text{hifi}}\times T}$ while preserving implicit spatial cues (e.g., IPD/ILD). First, we apply a shared weight-normalized 1D convolution to each channel independently to extract local features $\mathbf{X}_{\text{feat}} \in \mathbb{R}^{B\times C\times d\times T}$. To capture nonlinear inter-channel dependencies, we then employ Multi-Head Self-Attention along the channel axis at each time step. Unlike fixed difference operations, this data-driven approach dynamically weights the contribution of each channel. Finally, the attended features are concatenated and projected via a 1×1 convolution to the backbone dimension d_{hifi}, serving as the unified input to the generator.

3.4.2 Spatial Position Adaptor

This adaptor converts the raw pose sequence $\mathbf{P}$ into dense, physically meaningful conditioning $\mathbf{X}_{\text{pos}}$.

Feature Extraction & Encoding: From the 7D raw pose, we derive Cartesian coordinates and forward vectors (from quaternions), augmented with first-order velocity differences to capture kinematic motion. To mitigate the spectral bias of MLPs, we map these scalars to high-dimensional sinusoidal representations using Fourier feature encoding (Mildenhall et al., 2021), enabling sensitivity to fine-grained spatial changes.

Temporal Modeling & Injection: The encoded features are processed by CausalPosEncoder (stacked causal dilated convolutions) to model motion trajectories. We inject this condition into the generator via Feature-wise Linear Modulation (FiLM). For each upsampling block, audio features $\mathbf{x}_{\text{audio}}$ are modulated by scaling γ and bias β projected from the pose embeddings: $\text{FiLM}(\mathbf{x}_{\text{audio}}) = (1 + \tanh(\gamma)) \cdot \mathbf{x}_{\text{audio}} + \beta$.

3.5 Unified Framework for Spatial Audio

Traditional vocoders are mono-centric or naïvely replicate single-channel outputs, limiting their applicability to spatial audio. We design a channel-free generator where the shared backbone performs identical upsampling for any channel count: the Attentional Mel Adaptor fuses a C-channel mel-spectrogram into a fixed-dimensional representation, and the final projection layer outputs exactly C waveform channels. For adversarial training, we pair this flexible generator with channel-aware discriminator heads specialized for each format. At inference, a single checkpoint handles arbitrary supported formats by mapping the input mel-spectrogram and its channel configuration directly to spatial audio output. The design is naturally extensible: supporting new standards (e.g., 5.1 or 7.1 surround) requires only adding a format-specific spatial loss and discriminator head, without modifying the generator backbone.

4 Experiment

4.1 Experiment Details

Dataset We use both binaural and FOA formats data. For binaural data, we adopt the MRSSpeech subset of MRSAudio (Guo et al., 2025) and the EasyCom (Donley et al., 2021) dataset. For FOA data, we use the Spatial LibriSpeech (Sarabia et al., 2023) dataset, synthesized from LibriSpeech (Panayotov et al., 2015), which offers a large number of FOA samples with spatial annotations. To increase spatial and acoustic diversity, we further generate simulated data using the sound-space toolkit. In total, our training corpus contains roughly 600 hours of binaural data (about 350k samples) and 900 hours of FOA data (about 310k samples), all stored as 16-bit PCM at a sampling rate of 48 kHz.

We preprocess EasyCom and MRSSpeech datasets using the ClearVoice (Zhao et al., 2025b) denoising algorithm to enhance audio quality. We extract 700 random segments from all datasets as test set, then split the remaining dataset into training/validation sets with a ratio of 9:1. The detailed statistics are shown in appendix C.

Baseline We compare our proposed method with several vocoder baselines. We choose original HiFi-GAN (Kong et al., 2020), Vocos (Siuzdak, 2024) CARGAN (Morrison et al., 2022), FARGAN (Valin et al., 2024) and WaveFM (Luo et al., 2025) as our baselines. While recent works such as MusicHiFi (Zhu et al., 2024) have explored spatial audio vocoding, their implementations are not publicly available.Since there is a lack of dedicated spatial vocoder models for spatial audio generation, we select the above-mentioned baselines, which have demonstrated strong performance in monaural audio generation tasks. We perform channel-wise inference to generate binaural and FOA format audio for comparison with our model.

Metrics Our evaluation protocol comprises both subjective listening tests and objective metrics.

The objective evaluation addresses general audio quality and spectral/temporal similarity as well as spatial characteristics.

For waveform and spectral similarity we adopt the metrics used in BinauralGrad(Leng et al., 2022) MCD(Mel-cepstral distortion) to measure spectral distortion, Periodicity to assess periodicity in the audio. and MRSTFT, which combines spectral convergence with log- and linear-magnitude terms to improve spectral alignment. We also report PESQ as a perceptual measure for speech-related quality assessment. Except for PESQ, lower metric values indicate better performance.

To quantify spatial fidelity, we introduce two consistency measures ANG Cos and DIS Cos to respectively evaluate angular and distance similarity between generated and reference signals. Practically, we extract angular and distance embeddings from binaural audio using Spatial-AST. Because Spatial-AST(Zheng et al., 2024) produces position estimates only for static sources, we partition each audio into 1-second segments, compute the cosine similarity between predicted and ground-truth embeddings within each segment, and then average these segment-level similarities to obtain an overall spatial-consistency score. These metrics are report in percentage format.

we utilize subjective MOS-Q (Mean Opinion Score for Quality) to evaluate the quality of generated audio and MOS-P (Mean Opinion Score for Position) to assess spatial perception. Implementation details are in Appendix F.

4.2 Quantitative Comparison

We compare our model with existing vocoder baselines and present the metric results in Table 1. As shown in the table, our approach significantly outperforms all baselines on spatial metrics while achieving competitive results on audio metrics. This demonstrates that explicitly modeling inter-channel relationships through our Spatial Mel Adaptor and supervising spatial cues via the Spatial Consistency Discriminator are effective for preserving spatial information. For audio quality metrics, our model achieves qualitative reconstruction results. Our PESQ score is lower than non-causal SOTA baselines such as Vocos and WaveFM, which we attribute to the strictly causal constraint as our causal convolutions can only access past context, whereas non-causal models leverage bidirectional receptive fields that benefit perceptual quality. More results on FOA are in Appendix D.

We report the Real-Time Factor (RTF) measured on a single NVIDIA RTX 4090 GPU. Our model achieves RTF = 0.1587, which is well below unity and confirms that our causal streaming architecture supports real-time generation. And detailed results of latency experiment are in Appendix E.

These results demonstrate that our model effectively bridges the gap between high-fidelity waveform synthesis and accurate spatial rendering. The causal architecture introduces a minor quality trade-off compared to non-causal models, but this is an acceptable cost for enabling low-latency streaming applications.

4.3 Qualitative Comparison

We conduct a qualitative comparison of our proposed model with the baselines. We present the generated audio samples in Figure 3. The first row is the GT audio and the second is audio predicted by our model, followed by baseline predictions. Our causal model preserves the harmonic stacks and formant trajectories that closely match the ground truth on both channels, while maintaining consistent left-right spectral patterns. Compared with the baselines, our results exhibit sharper and more coherent harmonic structures with fewer band-wise artifacts and a cleaner noise floor. Although our causal generation still shows slightly smoother transients and mildly reduced high-frequency detail than non-causal counterparts, it achieves a highly similar overall spectral structure, indicating that high perceptual quality is attainable under causal constraints. We present more qualitative results in our demo page.

4.4 Subjective Evaluation

We conduct subjective listening tests to evaluate the quality and spatial perception of the generated audio.

We show the subjective evaluation result in Table 2. For spatial quality MOS-P test we ask listeners to rate how accurately they can perceive the position of the sound source in the generated audio compared to the ground truth position on a scale from 1 to 5. Our model achieves the highest MOS-P score among all models, indicating superior spatial perception. For audio quality MOS-Q test we ask listeners to rate the overall audio quality of the generated samples on a scale from 1 to 5. Our model also achieves high MOS-Q score, but causal generating may be slightly inferior in audio quality compared to non-causal models as they are able to utilize future context.

Model	**MOS-P**	**MOS-Q**
HiFi-GAN	3.86 ± 0.19	3.98 ± 0.17
CARGAN	3.90 ± 0.18	4.03 ± 0.14
FARGAN	3.93 ± 0.14	4.07 ± 0.15
WaveFM	4.13 ± 0.13	4.17 ± 0.12
Vocos	4.09 ± 0.15	4.24 ± 0.11
Ours	4.25 ± 0.16	4.09 ± 0.21
GT	4.42 ± 0.11	4.41 ± 0.16

Table 2: Subjective Evaluation Results

Model	**ANG COS** ($\uparrow$)	**DIS COS** ($\uparrow$)	**MRSTFT** ($\downarrow$)	**PESQ** ($\uparrow$)	**MCD** ($\downarrow$)	**Periodicity**($\downarrow$)	**RTF** ($\downarrow$)
HiFi-GAN	39.07	68.37	1.470	1.562	5.329	0.169	0.0622
CARGAN	30.00	63.71	1.194	1.739	3.377	0.160	0.1348
FARGAN	23.53	56.03	1.219	1.885	3.447	0.161	0.1916
WaveFM	41.36	71.96	1.079	2.400	2.727	0.141	0.1634
Vocos	40.04	70.23	**1.039**	**2.510**	**1.892**	0.113	**0.0339**
Ours	**62.11**	**77.05**	1.223	2.109	2.153	**0.107**	0.1587

Table 1: Quantitative Comparison

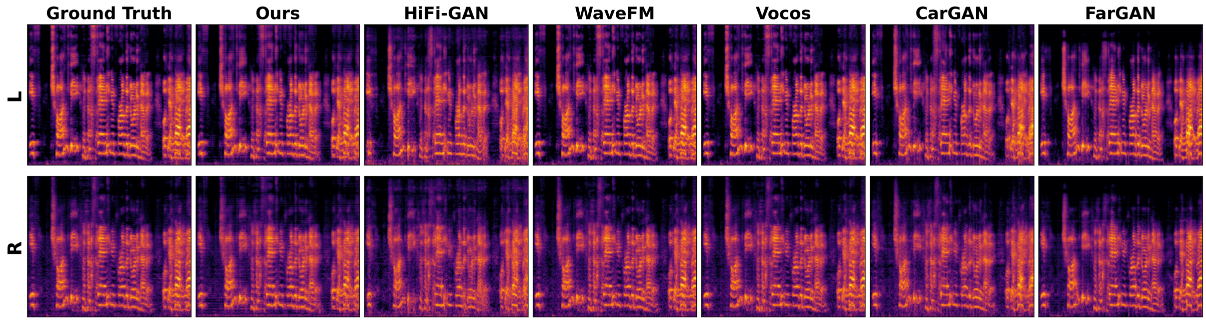

Figure 3: Qualitative comparison.

4.5 Ablation Study

We perform ablation studies of our proposed components and present the results in Table 3.

All proposed components contribute to the overall performance of our model. Removing Spatial Mel Adaptor causes significant drop in spatial metrics, as no inter-channel information is utilized. Using 4 attention heads in Spatial Mel Adaptor yields the best performance. However, removing the Position Adaptor results in moderate performance degradation, indicating that spatial information can still be partially captured through Spatial Mel Adaptor. The experiment shows that the Spatial Consistency Discriminator helps improve spatial metrics. But during our experiments, we find that only careful tuning of the weight of adversarial hyperparameters can lead to performance improvement, otherwise it may cause training instability.

Setting	**ANG COS** ($\uparrow$)	**DIS COS** ($\uparrow$)
w/o Mel Adaptor	42.60	65.39
Mel Adaptor 2 head	61.03	76.55
Mel Adaptor 8 head	61.50	76.70
w/o SCD	58.82	74.63
w/o Position Adaptor	54.78	70.63
Mel Adaptor 4 head	**62.11**	**77.05**

Table 3: Ablation Study

5 Conclusion

We present CSAVocoder, a spatial audio vocoder that jointly addresses high-fidelity waveform synthesis and accurate spatial rendering. Our framework extends the GAN architecture with three key innovations: (1) a spatial adaptor that fuses multi-channel mel-spectrograms with dynamic pose information to capture inter-channel relationships, (2) a spatial consistency discriminator that explicitly supervises spatial cues, and (3) a strictly causal, stateful generator that enables efficient streaming inference with constant memory overhead.

Experimental results demonstrate that CSAVocoder outperforms existing channel-wise vocoders in spatial fidelity and synthesis well audio quality while maintaining real-time performance. The universal architecture supports multiple spatial audio formats without format-specific modifications, making it a practical solution for immersive audio applications such as virtual reality, augmented reality, and spatial communication.

We hope that the explicit modeling of spatial information and the causal streaming design provide a strong foundation for future work on real-time spatial audio generation.

Limitations

Our work has three main limitations. First, establishing fair comparisons against causal baselines is challenging because different implementations adopt distinct buffering strategies and runtime optimizations that affect both latency and quality. Many strong vocoders are optimized for offline generation and benefit from non-causal context or heavier post-processing; even when adapted to streaming, their engineering choices can dominate measured runtime. A standardized causal-baseline suite with matched end-to-end latency budgets and consistent objective measurements is left for future work. Second, we focus on binaural and FOA formats; extending to higher-order ambisonics (HOA), multichannel loudspeaker layouts (e.g., 5.1/7.1), object-based audio, and personalized HRTF rendering is non-trivial. Increasing channel counts changes the required inductive bias, stability of adversarial training, and computational cost, and different ambisonic conventions may introduce dataset mismatches. Third, we condition on pose (position and orientation), but alternative or complementary representations may be more robust or expressive, such as relative geometry features (distance/azimuth/elevation), scene-aware embeddings from visual or 3D context, or learned spatial tokens that summarize multi-source environments. We do not exhaustively explore these design axes.

Ethical Considerations

This paper presents CSAVocoder, a causal and stateful vocoder for low-latency spatial audio generation conditioned on acoustic features. While the model does not generate linguistic content on its own, it can be integrated into upstream TTS/VC systems; therefore, both model- and data-related risks must be considered.

Data provenance, licensing, and privacy. We rely on publicly available speech/spatial-audio corpora and simulation pipelines. We do not claim ownership of any third-party audio content and recommend that any release avoid redistributing raw audio unless explicitly permitted by original licenses/terms. Derived artifacts such as file lists, splits, and evaluation scripts should be shared in a way that enables reproducibility while reducing privacy exposure. Speech datasets may contain personally identifying information or sensitive attributes.

Risks from real-time generation and speech privacy. Low-latency speech generation can enable near-real-time impersonation, "live" spoofing in voice authentication, and the re-synthesis of intercepted private conversations. Spatial audio further increases realism and may strengthen deceptive scenarios. In addition, pose conditioning introduces an auxiliary privacy surface: logged 3D trajectories and orientations can reveal behavioral patterns, attention, or activity context in immersive systems.

Potential harmful applications. Beyond deepfakes, potential misuse includes covert surveillance, harassment, social engineering, or generating misleading evidence. Dataset misuse may include training downstream models for speaker identification, demographic profiling, or other applications that participants did not consent to, especially when data is repurposed outside its original scope.

Mitigations and responsible release. We recommend (i) clear acceptable-use terms and licenses; (ii) optional watermarking/provenance signals and guidance for detection; (iii) restricting and documenting deployment contexts; (iv) minimizing retention of raw audio, intermediate representations, and pose logs; and (v) reporting limitations and failure modes. For listening tests, risks are minimal but include fatigue; conservative volume, breaks, and withdrawal options are advised.

Bias and environmental impact. Training data and simulators may under-represent languages, accents, acoustic environments, and accessibility-related speech characteristics, leading to uneven performance. Finally, while causal inference can reduce runtime cost, training remains compute-intensive; we encourage transparent reporting of compute and settings to support reproducibility and responsible scaling.

References

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.

Aryan Chaudhary and Vinayak Abrol. 2024. QGAN: Low Footprint Quaternion Neural Vocoder for Speech Synthesis. In *Interspeech 2024*, pages 3874–3878.

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. In *Advances in Neural Information Processing Systems*, volume 35, pages 8896–8911. Curran Associates, Inc.

Zhuangqi Chen, Xianjun Xia, Xiaohuai Le, Siyu Sun, and Chuanzeng Huang. 2025. AF-Vocoder: Artifact-Free Neural Vocoder with Global Artifact Filter. In *Interspeech 2025*, pages 4903–4907.

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*.

Hui-Peng Du, Yang Ai, and Zhen-Hua Ling. 2025. A distilled low-latency neural vocoder with explicit amplitude and phase prediction. *Preprint*, arXiv:2509.13667.

Ruohan Gao and Kristen Grauman. 2019. 2.5d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wenxiang Guo, Changhao Pan, Zhiyuan Zhu, Xintong Hu, Yu Zhang, Li Tang, Rui Yang, Han Wang, Zongbao Zhang, Yuhan Wang, and 1 others. 2025. Mrsaudio: A large-scale multimodal recorded spatial audio dataset with refined annotations. *arXiv preprint arXiv:2510.10396*.

Rishabh Gupta, Jianjun He, Rishabh Ranjan, Woon-Seng Gan, Florian Klein, Christian Schneiderwind, Annika Neidhardt, Karlheinz Brandenburg, and Vesa Välimäki. 2022. Augmented/mixed reality audio for hearables: Sensing, control, and rendering. *IEEE Signal Processing Magazine*, 39(3):63–89.

Mojtaba Heydari, Mehrez Souden, Bruno Conejo, and Joshua Atkins. 2025. Immersediffusion: A generative spatial audio latent diffusion model. In *ICASSP*.

Dhanush Joy, S Kiran, Alan Ponnachan, R Ashok, and M Nikhil. 2024. Real-time implementation of spatial audio on fpga using interaural time difference and amplitude-driven perceptual depth. In *2024 First International Conference on Electronics, Communication and Signal Processing (ICECSP)*, pages 1–6. IEEE.

Ganesh Kailas and Nachiketa Tiwari. 2021. Design for immersive experience: Role of spatial audio in extended reality applications. In *Design for Tomorrow—Volume 2: Proceedings of ICoRD 2021*, pages 853–863. Springer.

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

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. Diffwave: A versatile diffusion model for audio synthesis. *Preprint*, arXiv:2009.09761.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, and Aaron Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Preprint*, arXiv:1910.06711.

Luca A. Lanzendörfer, Florian Grötschla, Michael Ungersböck, and Roger Wattenhofer. 2025. High-fidelity music vocoder using neural audio codecs. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2022. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*.

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.

Susan Liang, Dejan Markovic, Israel D. Gebru, Steven Krenn, Todd Keebler, Jacob Sandakly, Frank Yu, Samuel Hassel, Chenliang Xu, and Alexander Richard. 2025. Binauralflow: A causal and streamable approach for high-quality binaural speech synthesis with flow matching models. *Preprint*, arXiv:2505.22865.

Xikun Lu, Yunda Chen, Zehua Chen, Jie Wang, Mingxing Liu, Hongmei Hu, Chengshi Zheng, Stefan Bleeck, and Jinqiu Sang. 2025. Deep learning for personalized binaural audio reproduction. *arXiv preprint arXiv:2509.00400*.

Tianze Luo, Xingchen Miao, and Wenbo Duan. 2025. Wavefm: A high-fidelity and efficient vocoder based on flow matching. *Preprint*, arXiv:2503.16689.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106.

Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio. 2022. Chunked autoregressive gan for conditional waveform synthesis. In *Submitted to ICLR 2022*.

Tan Dat Nguyen, Ji-Hoon Kim, Youngjoon Jang, Jaehun Kim, and Joon Son Chung. 2024. Fregrad: Lightweight and fast frequency-aware diffusion vocoder. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10736–10740.

Tianrui Pan, Jie Liu, Zewen Huang, Jie Tang, and Gangshan Wu. 2025. In-the-wild audio spatialization with flexible text-guided localization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1989–2001, Vienna, Austria. Association for Computational Linguistics.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Nikunj Raghuvanshi and John Snyder. 2018. Parametric directional coding for precomputed sound propagation. *ACM Transactions on Graphics (TOG)*, 37(4):1–14.

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*.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Saksham Singh Kushwaha, Jianbo Ma, Mark R. P. Thomas, Yapeng Tian, and Avery Bruni. 2024. Diff-SAGe: End-to-End Spatial Audio Generation Using Diffusion Models. *arXiv e-prints*, arXiv:2410.11299.

Hubert Siuzdak. 2024. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. *Preprint*, arXiv:2306.00814.

Christian J. Steinmetz and Joshua D. Reiss. 2020. auraloss: Audio focused loss functions in PyTorch. In *Digital Music Research Network One-day Workshop (DMRN+15)*.

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*.

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

Jean-Marc Valin, Ahmed Mustafa, and Jan Büthe. 2024. Very low complexity speech synthesis using framewise autoregressive gan (fargan) with pitch prediction. *IEEE Signal Processing Letters*, 31:2115–2119.

Simon Welker, Tal Peer, and Timo Gerkmann. 2025. Real-time streaming mel vocoding with generative flow matching. *Preprint*, arXiv:2509.15085.

Shahrokh Yadegari, John Burnett, Grady Kestler, and Louis Pisha. 2022. Spatial audio and sound design in the context of games and multimedia. In *Encyclopedia of Computer Graphics and Games*, pages 1–7. Springer.

Reo Yoneyama, Masaya Kawamura, Ryo Terashima, Ryuichi Yamamoto, and Tomoki Toda. 2025a. Comparative analysis of fast and high-fidelity neural vocoders for low-latency streaming synthesis in resource-constrained environments. *Preprint*, arXiv:2506.03554.

Reo Yoneyama, Atsushi Miyashita, Ryuichi Yamamoto, and Tomoki Toda. 2025b. Wavehax: Aliasing-free neural waveform synthesis based on 2d convolution and harmonic prior for reliable complex spectrogram estimation. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4454–4470.

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

Yu Zhang, Baotong Tian, and Zhiyao Duan. 2025b. Conan: A chunkwise online network for zero-shot adaptive voice conversion. *Preprint*, arXiv:2507.14534.

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

Shengkui Zhao, Zexu Pan, and Bin Ma. 2025b. Clearervoice-studio: Bridging advanced speech processing research and practical deployment. *arXiv preprint arXiv:2506.19398*.

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*.

Ge Zhu, Juan-Pablo Caceres, Zhiyao Duan, and Nicholas J Bryan. 2024. Musichifi: Fast high-fidelity stereo vocoding. *IEEE Signal Processing Letters*.

Zhiyuan Zhu, Yu Zhang, Wenxiang Guo, Changhao Pan, and Zhou Zhao. 2025. Asaudio: A survey of advanced spatial audio research. *arXiv preprint arXiv:2508.10924*.

A Implementation Details

This appendix provides the detailed hyperparameters and architectural configurations used in our experiments.

A.1 Audio and Spectrogram Parameters

All audio processing and mel-spectrogram extraction are conducted using the parameters listed in Table 4. The overall upsampling factor of the generator is set to 320 to match the hop size used in mel-spectrogram extraction.

Table 4: Audio processing and mel-spectrogram extraction parameters

Parameter	Value
Sample rate	48,000 Hz
FFT size	1024
Hop size	320
Window size	1024
Number of mel bins	128
Mel $f_{\min}$	20 Hz
Mel $f_{\max}$	24,000 Hz

A.2 Generator Architecture

The generator backbone G is based on the HiFi-GAN V1 configuration and is modified to support causal streaming synthesis. The total upsampling factor is $8 \times 5 \times 4 \times 2 = 320$. The detailed configuration is shown in Table 5.

A.3 Spatial adaptor Architecture

The spatial adaptor consists of two core submodules: the attention-based mel adaptor and the spatial position adaptor. Their configurations are summarized in Table 6.

A.4 Discriminator Configuration

We employ a combination of four discriminators to evaluate the generated audio from complementary perspectives. Table 7 summarizes their configurations.

A.5 Training and Optimization Hyperparameters

The training and optimization hyperparameters are listed in Table 8. We adopt a standard adversarial training setup with additional spectral and spatial losses.

B Losses Design

B.1 Adversarial Objective ($\mathcal{L}_{\text{adv}}$)

We adopt the Least-Squares GAN (LS-GAN) for adversarial training. For each discriminator D_k in the set $\{D_k\}$, the discriminator loss is $\mathcal{L}_{\text{adv}}(D_k, G) = \mathbb{E}_{\mathbf{y}}\big[(D_k(\mathbf{y}) - 1)^2\big] + \mathbb{E}_{\mathbf{z},\mathbf{c}}\big[D_k(G(\mathbf{z},\mathbf{c}))^2\big]$, which encourages high scores for real samples $\mathbf{y}$ and low scores for generated samples $G(\mathbf{z},\mathbf{c})$.

The generator adversarial loss is $\mathcal{L}_{\text{adv}}(G, D) = \sum_k \mathbb{E}_{\mathbf{z},\mathbf{c}}\big[(D_k(G(\mathbf{z},\mathbf{c})) - 1)^2\big]$, which encourages all discriminators to regard generated audio as real.

Since D comprises the MPD, MSD, MRD, and SCD introduced above, the final adversarial objectives are $\mathcal{L}_{\text{adv}}(G) = \sum_k \mathcal{L}_{\text{adv}}(G; D_k)$, $\mathcal{L}_{\text{adv}}(D) = \sum_k \mathcal{L}_{\text{adv}}(D_k; G)$, which jointly enforce alignment with real audio in temporal structure, multi-scale patterns, spectral detail, and spatial consistency.

B.2 Feature Matching Loss ($\mathcal{L}_{\text{fm}}$)

To stabilize GAN training and regularize the generator toward the real data manifold, we employ a feature matching loss. It acts as a perceptual constraint based on learned hierarchical representations: $\mathcal{L}_{\text{fm}}(G, D) = \sum_k \mathbb{E}_{\mathbf{y},\mathbf{z},\mathbf{c}}\left[\sum_{i=1}^{L_k} \frac{1}{N_i}\left\|D_k^{(i)}(\mathbf{y}) - D_k^{(i)}(G(\mathbf{z},\mathbf{c}))\right\|_1\right]$, where $D_k^{(i)}$ is the i-th intermediate feature map of discriminator D_k, L_k is the number of layers considered, and N_i is the number of elements in that feature map.

B.3 Auxiliary Perceptual and Reconstruction Losses

These losses provide more direct, non-adversarial gradient signals to the generator and optimize specific perceptual aspects of the synthesized audio.

To ensure that the spectral structure of the generated audio matches that of real audio, we employ two spectral reconstruction losses.

The first is the mel-spectrogram loss $\mathcal{L}_{\text{mel}}$, which computes the L1 distance between the mel-spectrograms of the generated audio $G(\mathbf{M},\mathbf{P})$ and the real audio $\mathbf{y}$. This loss constrains the model on the perceptually important mel scale and is defined as

$$\mathcal{L}_{\text{mel}}(G) = \mathbb{E}_{\mathbf{y},\mathbf{M},\mathbf{P}}\left[\left\|\phi(\mathbf{y}) - \phi(G(\mathbf{M},\mathbf{P}))\right\|_1\right], \tag{4}$$

where ϕ denotes the transformation from the waveform to its mel-spectrogram.

Layer / Block	Output Channels	Kernel	Stride	Upsample
Initial conv (`conv_pre`)	512	7	1	–
Upsampling block 1				
Causal upsampling	256	16	8	×8
MRF residual blocks	256	[3, 7, 11]	–	–
Upsampling block 2				
Causal upsampling	128	10	5	×5
MRF residual blocks	128	[3, 7, 11]	–	–
Upsampling block 3				
Causal upsampling	64	8	4	×4
MRF residual blocks	64	[3, 7, 11]	–	–
Upsampling block 4				
Causal upsampling	32	4	2	×2
MRF residual blocks	32	[3, 7, 11]	–	–
Final conv (`conv_post`)	C	7	1	–

Table 5: Generator backbone configuration

Submodule	Hyperparameter	Value
Attention Mel adaptor	Input mel bins	128
	Hidden channels	256
	Conv kernel size	5
	Number of attention heads	4
Spatial Position Adaptor	Input pose dimension	7
	Fourier feature bands	8
	Causal temporal encoder layers	3
	Temporal encoder kernel size	3
	Injection mechanism	FiLM
	Injection feature dimension	256

Table 6: Spatial adaptor configuration

The second is the multi-resolution STFT loss $\mathcal{L}_{\text{STFT}}$. This loss is computed under multiple short-time Fourier transform (STFT) configurations, each with different FFT sizes, window sizes, and hop sizes. It consists of two components: the spectral convergence loss $\mathcal{L}_{\text{sc}}$, which penalizes differences in spectral magnitude, and the log STFT magnitude loss $\mathcal{L}_{\text{mag}}$, which computes an L1 loss on the log-magnitude spectrogram and better reflects human perception of loudness. The total STFT loss is defined as the average of these two components across all STFT resolutions.

B.4 Spatial Loss Formulation

We provide the full formulation of the spatial loss $\mathcal{L}_{\text{spatial}}$, which explicitly supervises inter-channel spatial cues beyond per-channel spectral similarity. Its concrete form is defined in a format-adaptive way for binaural and First-Order Ambisonics (FOA) signals.

Binaural Spatial Loss. For binaural signals, we compute complex STFTs of the left and right channels, $S_L(f,t)$ and $S_R(f,t)$, under multiple STFT configurations. The interaural phase difference (IPD) is given by $\Delta\Phi(f,t) = \arg S_L(f,t) - \arg S_R(f,t)$. To avoid phase wrapping, we embed $\Delta\Phi$ into the complex plane and define

$$\mathbf{u}_{\text{IPD}}(f,t) = \big(\cos\Delta\Phi(f,t),\, \sin\Delta\Phi(f,t)\big) \in \mathbb{R}^2.$$

The IPD loss compares the embedded representations of the target and generated signals,

$$\mathcal{L}_{\text{IPD}} = \frac{\sum_{f,t} w_{\text{IPD}}(f)\, m(f,t) \left\| \mathbf{u}_{\text{IPD}}^{\text{pred}}(f,t) - \mathbf{u}_{\text{IPD}}^{\text{ref}}(f,t) \right\|_2^2}{\sum_{f,t} w_{\text{IPD}}(f)\, m(f,t) + \varepsilon},$$

Discriminator	Hyperparameter	Value
MPD	Periods	[2, 3, 5, 7, 11, 13, 17, 19, 23, 37]
MSD	Scales	raw, $\times 2$ pooling, $\times 4$ pooling
MRD	Resolution 1	[1024, 120, 600]
	Resolution 2	[2048, 240, 1200]
	Resolution 3	[512, 50, 240]
Attentional SCD	Backbone type	Axial attention
	Number of axial attention blocks	2
	Number of attention heads	4

Table 7: Discriminator configuration, MRD resolutions are specified as [FFT size, hop size, window size]

Hyperparameter	Value
Optimizer	Adam
Learning rate (G / D)	2×10^{-4}
Adam betas (β_1, β_2)	(0.8, 0.99)
Learning rate decay γ	0.999
Batch size	16
Audio segment length	16,384 samples
Loss weights	
$\mathcal{L}_{\text{adv}}$ (λ_{adv})	1.0
$\mathcal{L}_{\text{fm}}$ (λ_{fm})	2.0
$\mathcal{L}_{\text{mel}}$ (λ_{mel})	45.0
$\mathcal{L}_{\text{STFT}}$ (λ_{STFT})	1.0
$\mathcal{L}_{\text{spatial}}$ (IPD/ILD)	0.1
$\mathcal{L}_{\text{spatial}}$ (FOA)	2.0

Table 8: Training and optimization hyperparameters

where $w_{\text{IPD}}(f) = \exp\big(-(f/f_{\text{IPD,max}})^2\big)$ emphasizes low frequencies and $m(f,t)$ is an energy-based soft mask.

The interaural level difference (ILD) is defined in the log-magnitude domain as

$$\text{ILD}^{\text{ref}}(f,t) = 20\log_{10}|S_L^{\text{ref}}(f,t)| - 20\log_{10}|S_R^{\text{ref}}(f,t)|,$$

and analogously for ILD^{pred}. The ILD loss is

$$\mathcal{L}_{\text{ILD}} = \frac{\sum_{f,t} w_{\text{ILD}}(f)\, m(f,t)\, \big|\text{ILD}^{\text{pred}}(f,t) - \text{ILD}^{\text{ref}}(f,t)\big|}{\sum_{f,t} w_{\text{ILD}}(f)\, m(f,t) + \varepsilon},$$

with $w_{\text{ILD}}(f) = 1 - \exp\big(-(f/f_{\text{ILD,min}})^2\big)$ that emphasizes high frequencies.

The soft mask $m(f,t)$ is derived from the frame-wise energy of the reference signal. Let $E(t)$ be the RMS energy at frame t (averaged over frequency and channels), and

$$E_{\text{dB}}(t) = 10\log_{10}(E(t) + \varepsilon).$$

We define a smooth frame-wise speech activity

$$s(t) = \sigma\left(\frac{E_{\text{dB}}(t) - \mu_{\text{VAD}}}{\sigma_{\text{VAD}}}\right),$$

where $\sigma(\cdot)$ is the sigmoid function, μ_{VAD} is the soft-VAD center in dB, and σ_{VAD} controls the transition width. The time–frequency mask is then

$$m(f,t) = m_{\min} + (1 - m_{\min})\, s(t),$$

with $m_{\min} > 0$ to avoid nullifying silent regions. The binaural spatial loss is

$$\mathcal{L}_{\text{spatial}}^{\text{bin}} = \lambda_{\text{IPD}}\mathcal{L}_{\text{IPD}} + \lambda_{\text{ILD}}\mathcal{L}_{\text{ILD}}.$$

FOA Spatial Loss. For FOA signals, we assume a B-format ordering (W, X, Y, Z). Given target and predicted waveforms $y, \hat{y} \in \mathbb{R}^{B\times 4\times T}$, we compute complex STFTs for each scale, obtaining

$$W(f,t),\ X(f,t),\ Y(f,t),\ Z(f,t)$$

for the reference and $\hat{W}(f,t),\ \hat{X}(f,t),\ \hat{Y}(f,t),\ \hat{Z}(f,t)$ for the prediction. The total FOA energy at each time–frequency bin is

$$E^{\text{ref}}(f,t) = |W|^2 + |X|^2 + |Y|^2 + |Z|^2,$$
$$E^{\text{pred}}(f,t) = |\hat{W}|^2 + |\hat{X}|^2 + |\hat{Y}|^2 + |\hat{Z}|^2.$$

Energy-weighted mask and frequency biases. We reuse the soft mask $m(f,t)$ from the binaural case, now interpreted per FOA STFT configuration. To steer supervision across frequency, we define a low-frequency bias for direction-related terms,

$$w_{\text{dir}}(f) = \exp\big(-(f/f_{\text{iv,max}})^2\big),$$

and a smooth mid–high-frequency bias for diffuseness-related terms. Let f_s be the sampling rate and $\tilde{f} = f/(f_s/2)$ the normalized frequency. We set

$$w_{\text{diff}}(f) = \tfrac{1}{2} + \tfrac{1}{2}\tanh\left(\frac{\tilde{f} - c_{\text{diff}}}{w_{\text{diff}}}\right),$$

where c_{diff} controls the center of the transition and w_{diff} controls its width.

We also apply mild energy exponents E^{α} to emphasize high-energy bins without dominating the loss. We denote these exponents by $\alpha_{\text{iv}}, \alpha_{\text{r}}, \alpha_{\text{diff}}$.

Intensity vector and directional term. The active intensity components are computed as

$$I_X^{\text{ref}}(f,t) = \Re\{W^*(f,t)X(f,t)\},$$
$$I_Y^{\text{ref}}(f,t) = \Re\{W^*(f,t)Y(f,t)\},$$
$$I_Z^{\text{ref}}(f,t) = \Re\{W^*(f,t)Z(f,t)\},$$

and analogously for $I_X^{\text{pred}}, I_Y^{\text{pred}}, I_Z^{\text{pred}}$. We collect these into intensity vectors

$$\mathbf{I}^{\text{ref}}(f,t) = [I_X^{\text{ref}}, I_Y^{\text{ref}}, I_Z^{\text{ref}}]^{\top},$$
$$\mathbf{I}^{\text{pred}}(f,t) = [I_X^{\text{pred}}, I_Y^{\text{pred}}, I_Z^{\text{pred}}]^{\top}.$$

The directional mismatch is measured via the cosine distance

$$d_{\text{iv}}(f,t) = 1 - \frac{\mathbf{I}^{\text{ref}}(f,t)^{\top}\mathbf{I}^{\text{pred}}(f,t)}{\|\mathbf{I}^{\text{ref}}(f,t)\|_2\,\|\mathbf{I}^{\text{pred}}(f,t)\|_2 + \varepsilon},$$

and we define

$$\mathcal{L}_{\text{iv\_dir}} = \frac{\sum_{f,t} m(f,t)\, w_{\text{dir}}(f)\, \left(E^{\text{ref}}(f,t)\right)^{\alpha_{\text{iv}}} d_{\text{iv}}(f,t)}{\sum_{f,t} m(f,t)\, w_{\text{dir}}(f)\, \left(E^{\text{ref}}(f,t)\right)^{\alpha_{\text{iv}}} + \varepsilon}.$$

Normalized intensity ratio term. We normalize the intensity by total energy,

$$\mathbf{r}^{\text{ref}}(f,t) = \frac{\mathbf{I}^{\text{ref}}(f,t)}{E^{\text{ref}}(f,t) + \varepsilon},$$
$$\mathbf{r}^{\text{pred}}(f,t) = \frac{\mathbf{I}^{\text{pred}}(f,t)}{E^{\text{pred}}(f,t) + \varepsilon},$$

and define

$$\mathcal{L}_{\text{r}} = \frac{\sum_{f,t} m(f,t)\, w_{\text{dir}}(f)\, \left(E^{\text{ref}}(f,t)\right)^{\alpha_{\text{r}}} \left\|\mathbf{r}^{\text{pred}}(f,t) - \mathbf{r}^{\text{ref}}(f,t)\right\|_1}{\sum_{f,t} m(f,t)\, w_{\text{dir}}(f)\, \left(E^{\text{ref}}(f,t)\right)^{\alpha_{\text{r}}} + \varepsilon}.$$

Diffuseness term. We compute the intensity norm

$$\|\mathbf{I}^{\text{ref}}(f,t)\|_2, \quad \|\mathbf{I}^{\text{pred}}(f,t)\|_2,$$

and define diffuseness as

$$D^{\text{ref}}(f,t) = 1 - \frac{\|\mathbf{I}^{\text{ref}}(f,t)\|_2}{E^{\text{ref}}(f,t) + \varepsilon},$$
$$D^{\text{pred}}(f,t) = 1 - \frac{\|\mathbf{I}^{\text{pred}}(f,t)\|_2}{E^{\text{pred}}(f,t) + \varepsilon}.$$

The diffuseness loss is then

$$\mathcal{L}_{\text{diff}} = \frac{\sum_{f,t} m(f,t)\, w_{\text{diff}}(f)\, \left(E^{\text{ref}}(f,t)\right)^{\alpha_{\text{diff}}} \left(D^{\text{pred}}(f,t) - D^{\text{ref}}(f,t)\right)^2}{\sum_{f,t} m(f,t)\, w_{\text{diff}}(f)\, \left(E^{\text{ref}}(f,t)\right)^{\alpha_{\text{diff}}} + \varepsilon}.$$

Log-energy term. Finally, we align the log-energy fields of reference and prediction:

$$\log E^{\text{ref}}(f,t) = \log(E^{\text{ref}}(f,t) + \varepsilon),$$
$$\log E^{\text{pred}}(f,t) = \log(E^{\text{pred}}(f,t) + \varepsilon),$$

and define

$$\mathcal{L}_{\text{elog}} = \frac{\sum_{f,t} m(f,t)\, w_{\text{diff}}(f) \left|\log E^{\text{pred}}(f,t) - \log E^{\text{ref}}(f,t)\right|}{\sum_{f,t} m(f,t)\, w_{\text{diff}}(f) + \varepsilon}$$

Multi-scale aggregation. In practice, all the above quantities are computed for multiple STFT parameter sets $(n_{\text{FFT}}, \text{hop}, \text{win})$. The four FOA terms $\mathcal{L}_{\text{iv\_dir}}, \mathcal{L}_{\text{r}}, \mathcal{L}_{\text{diff}}, \mathcal{L}_{\text{elog}}$ are averaged over scales, and the final FOA spatial loss is

$$\mathcal{L}_{\text{spatial}}^{\text{FOA}} = \lambda_{\text{iv}}\mathcal{L}_{\text{iv\_dir}} + \lambda_{\text{r}}\mathcal{L}_{\text{r}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{elog}}\mathcal{L}_{\text{elog}}.$$

C Details of Datasets

C.1 Recorded Binaural and FOA Data

We use both binaural and first-order Ambisonics (FOA) spatial audio data for training and evaluation. For the binaural branch, we adopt the MRSSpeech subset of the MRSAudio (Guo et al., 2025) corpus together with the EasyCom (Donley et al., 2021) dataset, which contain extensive indoor recordings captured with binaural microphones. These corpora cover multiple speakers, diverse source-listener spatial configurations, and both Chinese and English speech, providing realistic binaural characteristics and room acoustics. For FOA, we use the Spatial LibriSpeech (Sarabia et al., 2023) dataset, which is synthesized from LibriSpeech (Panayotov et al., 2015) and provides a large number of FOA-format spatial speech samples with corresponding position annotations. However, Spatial LibriSpeech only models azimuthal variation on the horizontal plane during spatialization and lacks diversity along the vertical dimension (elevation). This may result in that most samples would show silence in the z channel, potentially causing the model to overfit spatial perception on the horizontal plane while lacking sensitivity to the vertical direction.

C.2 Simulated Spatial Data from SoundSpaces (MP3D)

To enrich spatial diversity, especially in elevation and in complex 3D room geometries, we additionally generate a large amount of simulated spatial data based on the SoundSpaces (Chen et al., 2022) simulation framework and Habitat-Sim (Savva et al., 2019). In this work we focus on indoor scenes from the Matterport3D (MP3D) dataset; for each MP3D environment we instantiate a Habitat-Sim simulator and attach an audio sensor configured either as binaural (2-channel) or FOA Ambisonics (4-channel) at a sampling rate of 48 kHz. The listener (receiver) is placed at a height of 1.5 m above the floor, and the audio materials configuration from MP3D is loaded to enable frequency-dependent reflection, absorption, and diffraction in the propagation engine. We calculate the relative pose between source and receiver, and use it as conditioning input to the model.

C.3 Static BRIR/RIR Sampling and Position Generation.

For the static subset, we randomly sample receiver and source positions on the MP3D navigation mesh. A candidate pair is accepted only if the horizontal distance lies within $(1, 10)$ m and the height difference is smaller than 2 m, which avoids degenerate configurations (too close or too far, or across floors). For each accepted pair we query the audio sensor once and obtain a binaural or FOA room impulse response (BRIR/RIR). All positions are initially given in the Habitat/BAT coordinate convention, where the horizontal plane is x-z, y points upwards, and the agent faces the $-z$ direction. For downstream usage we convert all 3D positions (x, y, z) into a more conventional, listener-centric coordinate system with the horizontal plane being x-y, z pointing upwards, and the listener facing the $+y$ direction. All relative positions (source minus receiver) stored in our dataset are expressed in this transformed coordinate system.

C.4 Dynamic Simulated Trajectories.

Besides the purely static BRIRs, we also construct a dynamic subset in which the listener remains fixed while the source moves through the environment. Concretely, for a given receiver position we randomly sample two source points that are both within a reasonable distance from the receiver and compute the shortest path between them on the navigation mesh. The resulting 3D path is uniformly subsampled to a fixed number of time steps (e.g., 20 frames per trajectory). At each step we update the source position in Habitat-Sim, query a new BRIR from the audio sensor, and record the corresponding source position, relative position, and coarse direction labels (`left/right`, `front/behind`, `above/below`) derived from the transformed coordinate system. For each utterance we also generate a frame-level pose sequence at 20 Hz by repeating the (static) relative position or by aligning it with the dynamic trajectory, yielding an $N \times 7$ pose matrix per audio sample that is fully time-synchronized with the waveform.

C.5 Convolution with Mono Speech and Post-processing.

To turn the simulated BRIR/RIRs into training data, we convolve them with clean, single-channel speech from the LibriSpeech (Panayotov et al., 2015) corpus. All LibriSpeech utterances are first resampled to 48kHz and converted to mono. For each utterance we randomly select one BRIR entry, perform FFT-based convolution to obtain either 2-channel binaural or 4-channel FOA audio, and then truncate the result to match the original utterance length. We apply simple peak normalization (with a conservative safety margin) to avoid clipping and ensure that all simulated samples are loudness-consistent with the real-world data.

C.6 Overall Dataset Scale

Combining the real and simulated corpora, our final training and evaluation set comprises approximately 600 hours of binaural data and 900 hours of FOA data. Among them, around 220k binaural samples and 70k FOA samples are synthesized by convolving LibriSpeech with SoundSpaces-generated BRIR/RIRs in MP3D environments, while the remaining samples come from MRSSpeech, EasyCom, and Spatial LibriSpeech. All audio is uniformly resampled to 48kHz, and all spatial annotations are provided in the unified listener-centric coordinate system.

D FOA Results

We present additional experimental results for FOA spatial audio synthesis. For FOA audio, we adopt similar evaluation metrics as for binaural audio, including audio quality metrics (PESQ, MRSTFT, MCD) and spatial consistency metrics (Corr_all and AUC_j_all). The spatial consistency metrics

Model	Corr_all (↑)	AUC_ j _all (↑)	MRSTFT (↓)	MCD (dB) (↓)	PESQ (↑)
HiFi-GAN	18.65	61.98	1.278	4.052	2.122
CARGAN	15.99	61.20	1.257	3.690	1.757
FARGAN	16.26	61.28	1.154	2.941	1.794
WaveFM	14.37	60.80	0.846	1.950	3.520
Vocos	19.49	62.92	0.918	1.453	2.997
Ours	18.53	63.44	1.248	3.449	1.972

Table 9: FOA Results

are derived from the ViSAGe work (Kim et al., 2025) and assess the ability of the generated audio to preserve spatial cues. For audio quality, we use common metrics such as PESQ, MRSTFT, and MCD to measure the quality of the generated audio. For the FOA format, we specifically evaluate the audio quality of the W channel. Table 9 presents a quantitative comparison of our method against several baseline models on the FOA spatial audio synthesis task. It can be observed that our method outperforms others in spatial consistency metrics (Corr_all and AUC_j_all), indicating better performance in preserving spatial cues. Furthermore, our method achieves audio quality metrics comparable to non-causal models, demonstrating strong audio synthesis capabilities.

E Latency Evaluation

This appendix details how we define and measure latency for streaming inference, and reports representative results under different chunk sizes.

E.1 Definitions

For streaming audio generation, we consider three types of latency:

Algorithmic latency (L_{alg}, ms). This is the inherent delay introduced by the streaming design, independent of hardware speed. Under chunked inference, a system that outputs audio only after receiving a full chunk has a lower bound

$$L_{\text{alg}} \geq T_{\text{chunk}} + T_{\text{lookahead}} + T_{\text{overlap}}, \tag{5}$$

where T_{chunk} is the chunk duration, $T_{\text{lookahead}}$ is any future-context requirement (0 for strictly causal designs), and T_{overlap} accounts for cross-fade/overlap-add schemes that require waiting for future samples. For our model, $T_{\text{lookahead}} = 0$ and $T_{\text{overlap}} = 0$.

Compute latency (L_{comp}, ms/chunk). This is the wall-clock time to run the model for one chunk (forward pass in streaming mode). We report distributional statistics (p50/p90/p99) because tail latency is critical for real-time playback stability.

Real-Time Factor (RTF). To normalize compute latency across chunk sizes, we report

$$\text{RTF} = \frac{L_{\text{comp}}}{T_{\text{chunk}}}. \tag{6}$$

RTF < 1 indicates faster-than-real-time inference.

E.2 Chunking under sr $= 48$ kHz, hop=320

With sampling rate sr $= 48$ kHz and hop size 320 samples, the feature frame rate is

$$f = \frac{48000}{320} = 150 \text{ frames/s}, \tag{7}$$

Therefore, chunk sizes of 40/60/80/100 ms correspond to 6/9/12/15 mel frames, respectively.

E.3 Measurement protocol

We benchmark streaming inference with batch size 1 and disable gradient computation. For GPU timing, we synchronize before and after each forward pass to measure true kernel execution time. We perform a warm-up phase to avoid one-time compilation and cache effects, then run a fixed number of iterations and collect per-chunk latency samples, from which we compute mean and percentiles (p50/p90/p99).

E.4 Results and discussion

Table 10 reports representative compute latency under different chunk sizes. Across repeated runs, the mean compute latency stays in a narrow band (approximately 15ms/chunk), while RTF improves as chunk size increases. This behavior is expected on GPUs when sequence lengths are short: fixed overheads (kernel launches, framework scheduling, memory movements) can dominate, and larger chunks may better utilize the GPU, reducing the *per-frame* cost even if ms/chunk is similar. Importantly, all tested settings achieve RTF < 1, indicating real-time feasibility with substantial headroom.

Chunk	Mean	p50	p90	p99	RTF
40	15.24 ± 0.95	14.99	16.44	18.62	0.3811
60	15.15 ± 1.35	14.80	16.70	19.34	0.2526
80	15.52 ± 2.06	14.71	17.71	24.67	0.1941
100	15.86 ± 2.40	15.46	18.14	24.81	0.1587

Table 10: Representative compute latency (ms/chunk) for streaming inference at different chunk sizes under sr $= 48$ kHz and hop=320. We report p50/p90/p99 and RTF = $L_{\text{comp}}/T_{\text{chunk}}$.

F Details of Experiments

F.1 Subjective evaluation

The subjective evaluation is conducted in a controlled acoustic environment featuring sound-attenuated conditions, precisely calibrated playback systems, and frequency-equalized headphones to ensure consistency across listening sessions. A total of 200 audio segments are randomly sampled from the test dataset for evaluation purposes. We recruit 29 participants to provide perceptual ratings across two dimensions: audio quality and spatial perception, using a 5-point Likert scale ranging from 1 (Poor) to 5 (Excellent).

For audio quality assessment, we employ the Mean Opinion Score for Quality (MOS-Q), wherein participants utilize headphones to evaluate the clarity and naturalness of the synthesized audio. For spatial perception assessment, we adopt the Mean Opinion Score for Spatialization (MOS-P), where participants judge the authenticity of spatial attributes, including the correspondence between the perceived sound source localization (direction and distance) and the textual prompt specifications.

All participants receive appropriate compensation at an hourly rate of \$20, yielding a total experimental cost of approximately \$1500. Prior to participation, subjects are informed that their assessments will be utilized exclusively for academic research purposes. Detailed instructions provided to participants for the audio evaluation protocol are illustrated in Figure 4 and 5.

F.2 Objective evaluation

To ensure the reproducibility of our experiments, we employ standard open-source implementations for objective evaluation. The specific configurations and libraries used are detailed below:

MRSTFT: We utilize the Multi-Resolution Short-Time Fourier Transform (MRSTFT) implementation from Auraloss (Steinmetz and Reiss, 2020). The metric is computed as the sum of spectral convergence and log-magnitude distance across multiple window sizes.
`https://github.com/csteinmetz1/auraloss`

PESQ: Perceptual Evaluation of Speech Quality (PESQ) is evaluated using the Wideband mode (ITU-T P.862.2). Since our model generates 48 kHz audio, we downsample both the reference and synthesized signals to 16 kHz solely for this measurement using the `python-pesq` wrapper.
`https://github.com/ludlows/python-pesq`

MCD: We compute the Mel-Cepstral Distortion (MCD) to measure the spectral envelope difference. We use the `mel-cepstral-distance` library with Dynamic Time Warping (DTW) enabled to align the sequences before calculation.
`https://github.com/MattShannon/mcd`

Periodicity: To evaluate pitch accuracy and harmonic consistency, we calculate the periodicity error using the pre-trained CREPE model provided in the CARGAN repository (Morrison et al., 2022). The metric represents the root mean squared error between the periodicity vectors of the ground truth and generated audio.
`https://github.com/descriptinc/cargan`

ANG COS & DIS COS: To quantify spatial fidelity, we utilize the pre-trained Spatial-AST model (Zheng et al., 2024) to extract high-level spatial representations. We report the metrics as ANG COS (for angular consistency) and DIS COS (for distance consistency), where higher cosine similarity indicates better preservation of perceptible spatial cues.
`https://github.com/zszheng147/Spatial-AST`

RTF: Real-Time Factor (RTF) is calculated as the time required to generate the waveform divided by the duration of the audio on a single NVIDIA 4090 GPU.

G Licenses and Availability

We respect the original licenses of all referenced artifacts and do not redistribute them. This work uses publicly available datasets. We do not redistribute any third-party audio content. Users must obtain the original datasets from their respective providers and comply with the original licenses/terms of use. We will release only derived metadata (e.g., file lists, splits, and non-invertible statistics) under CC

Figure 4: This is a screenshot of our MOS-P test website

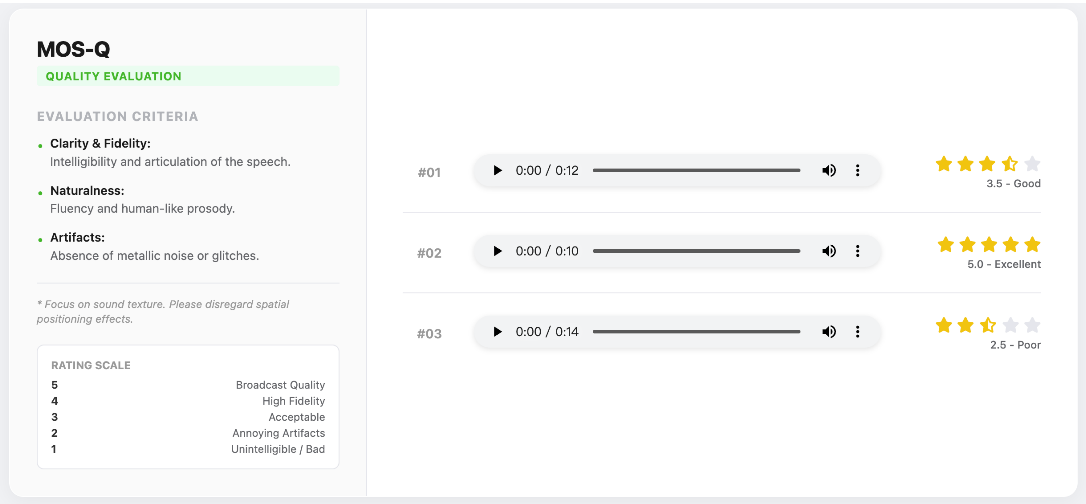

Figure 5: This is a screenshot of our MOS-Q test website

BY 4.0, subject to the original dataset terms. Our codebase may depend on third-party libraries; these components remain under their respective licenses. Any external assets (e.g., pretrained backbones or evaluation tools) are used in accordance with their original licensing terms.

H Use of AI Assistants

We used AI-based writing assistant during manuscript preparation solely for language polishing, including grammar checking, spelling correction, and improving clarity and readability of the text. All technical claims, experimental procedures, and interpretations were produced and verified by the authors.