SpatialV2A: Visual-Guided High-fidelity Spatial Audio Generation

Yanan Wang[1] , **Linjie Ren**[1] , **Zihao Li**[1] , **Junyi Wang**[1] , **Tian Gan**[1,*]
[1] Shandong University
wangyanan99@sdu.edu.cn, {renlinjie,lizihao20011219}@mail.sdu.edu.cn,
{junyiwang,gantian}@sdu.edu.cn

Abstract

While video-to-audio generation has achieved remarkable progress in semantic and temporal alignment, most existing studies focus solely on these aspects, paying limited attention to the spatial perception and immersive quality of the synthesized audio. This limitation stems largely from current models' reliance on mono audio datasets, which lack the binaural spatial information needed to learn visual-to-spatial audio mappings. To address this gap, we introduce two key contributions: we construct BinauralVGGSound, the first large-scale video-binaural audio dataset designed to support spatially aware video-to-audio generation; and we propose an end-to-end spatial audio generation framework guided by visual cues that explicitly models spatial features. Our framework incorporates a visual-guided audio spatialization module that ensures the generated audio exhibits realistic spatial attributes and layered spatial depth while maintaining semantic and temporal alignment. Experiments show that our approach substantially outperforms state-of-the-art models in spatial fidelity and delivers a more immersive auditory experience, without sacrificing temporal or semantic consistency. All datasets, code, and model checkpoints are available at: https://github.com/renlinjie868-web/SpatialV2A.

1 Introduction

Video-to-Audio (V2A) generation aims to synthesize high-fidelity, perceptually consistent audio signals that accurately match visual inputs. The core objective is to produce audio that is semantically aligned with visual events, temporally synchronized with scene dynamics, and perceptually realistic for an immersive user experience [Dagli *et al.*, 2025; Jeong *et al.*, 2025; Wang *et al.*, 2024; Xing *et al.*, 2024]. By establishing robust cross-modal associations between visual cues and their corresponding acoustic patterns, V2A systems can generate audio that accurately reflects object interactions, environmental context, and temporal motion cues. Compared to manual Foley design, V2A offers higher production efficiency, lower labor costs, and scalability to large-scale video production [Huang *et al.*, 2025]. As a result, V2A technology is increasingly important for applications such as video generation, virtual and augmented reality, and AI-assisted media production, where high-quality, visually grounded audio plays a crucial role in enhancing perceptual engagement and overall user experience.

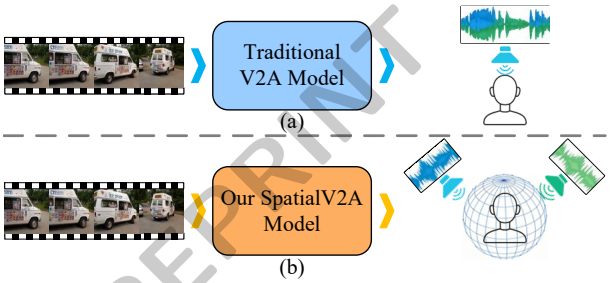

Figure 1: Comparison of traditional V2A models and the proposed SpatialV2A. (a) Traditional V2A generates mono audio. (b) SpatialV2A produces spatially consistent binaural audio.

Although recent V2A studies have achieved remarkable progress in generating semantically and temporally aligned audio, they have largely overlooked audio spatialization [Jeong *et al.*, 2025; Wang *et al.*, 2024; Xing *et al.*, 2024]. As illustrated in Figure 1(a), traditional V2A frameworks typically generate mono audio, which lacks directional and positional sound cues, thus failing to convey realistic spatial perception. However, spatial audio plays a crucial role in human auditory perception by providing cues about sound source direction, position, and movement, which are essential for immersive and realistic audiovisual experiences [Pan *et al.*, 2024]. Recent studies have begun to explore visual-guided audio spatialization, aiming to generate binaural or stereo audio from visual inputs [Dagli *et al.*, 2025]. Although these methods can synthesize spatialized audio, they often exhibit limited temporal and semantic alignment with visual events. For example, See-2-Sound [Dagli *et al.*, 2025] generates spatial audio solely from visual content without leveraging the real audio signal, which may lead to inconsistencies between the synthesized sound and the visual scene. Consequently, there is a critical need for a V2A framework that preserves

*Corresponding author.

Dataset	Type	Clips	Duration	Class	AV-C	Source	Caption
Fair-Play [Gao and Grauman, 2019]	Binaural	1.9k	5.2h	–	✓	Indoor recording	–
SimBinaural [Garg *et al.*, 2023]	Binaural	22k	116.1h	11	✓	Simulation	–
YouTube-Binaural [Garg *et al.*, 2023]	Binaural	0.4k	27h	–	–	YouTube	–
VGGSound [Chen *et al.*, 2020]	Mono	199k	550h	309	✓	YouTube	✓
BinauralVGGSound (ours)	Binaural	187k	519h	309	✓	YouTube	✓

Table 1: Comparison of existing audio-visual datasets. Many binaural datasets are small in scale, contain few categories, or lack captions, limiting their applicability to general V2A research. AV-C represents whether the audio signal and video scene are consistent.

semantic fidelity, temporal coherence, and spatial realism simultaneously, as illustrated by our proposed SpatialV2A in Figure 1(b).

Developing a spatially aware V2A framework faces two major challenges. First, achieving spatial perception in V2A generation requires an explicit audio spatialization module that can deduce spatial sound cues directly from visual scenes. Such a module must localize and temporally align sound-emitting events within the visual domain while capturing their relative intensity, thereby supplying the spatial information necessary for stereo-aware audio synthesis. Second, the development of spatial V2A models is hindered by the limited availability of large-scale binaural datasets. Real-world binaural recordings require specialized equipment and substantial human effort, making large-scale acquisition costly and impractical.

As shown in Table 1, most existing visual–audio datasets are inadequate for training spatially aware V2A models: some provide only mono audio (*e.g.*, VGGSound), while others are limited in scale, category diversity, or lack textual captions (*e.g.*, Fair-Play and SimBinaural). These limitations underscore the need for a large-scale and diverse dataset that can leverage existing visual information to convert mono audio into a binaural format—thereby enabling the construction of extensive stereo visual–audio pairs to advance spatial V2A research.

To address these limitations, this work makes the following contributions:

1. We construct a large-scale spatial video–audio dataset, **BinauralVGGSound**, comprising more than 187k high-quality video–audio pairs with binaural audio and descriptive captions. This dataset fills the gap in large-scale binaural-video datasets for V2A research, and establishes a data foundation for the training, inference, and advancement of spatially aware audio generation models.
2. We propose a spatially aware V2A generation framework **SpatialV2A** that models fine-grained spatial representations from visual inputs and synthesizes binaural stereo audio. Building upon the proposed dataset, our framework effectively integrates spatial perception into the V2A generation pipeline.
3. Extensive experiments demonstrate that our dataset construction strategy maintains the fidelity of the original mono audio, while the proposed framework generates binaural audio with precise temporal and semantic alignment. The results validate the effectiveness of both the dataset and the SpatialV2A framework for immersive audio-visual synthesis.

2 Related Work

2.1 Video-to-Audio

V2A generation aims to synthesize high-fidelity audio that is both semantically and temporally consistent with the visual content [Jeong *et al.*, 2025; Luo *et al.*, 2023]. Leveraging the availability of pretrained multimodal models, recent studies have made substantial progress in improving both *semantic* and *temporal* alignment in V2A generation.

Semantic Alignment. Semantic alignment ensures that the generated audio corresponds to the objects, actions, and scenes present in the video. Most existing V2A approaches achieve this by utilizing large-scale pretrained multimodal encoders. For instance, Im2Wav [Sheffer and Adi, 2023] and MMAudio [Jeong *et al.*, 2025] employ the CLIP encoder [Radford *et al.*, 2021] to align visual and textual semantics, while Diff-Foley [Luo *et al.*, 2023] introduces a contrastive audio-visual pretraining model to reinforce cross-modal correspondence. These models provide a strong semantic foundation, which our method adopts to maintain consistent semantic alignment throughout the generation process.

Temporal Alignment. Beyond semantic relevance, precise temporal alignment between visual events and audio signals is essential. Recent works address this through various temporal feature synchronization mechanisms. For example, SpecVQGAN [Iashin and Rahtu, 2021] conditions discrete audio representations on visual tokens to maintain temporal continuity. Diff-Foley [Luo *et al.*, 2023] leverages contrastive audio-visual pretraining to learn temporally aligned cross-modal embeddings. Furthermore, Synchformer [Iashin *et al.*, 2024] improves fine-grained synchronization via attention-based cross-modal alignment. Its effectiveness has been validated in subsequent works such as V-AURA [Viertola *et al.*, 2025], MMAudio [Cheng *et al.*, 2025], and ViSAudio [Zhang *et al.*, 2025], where integrating Synchformer significantly enhances audio–visual event alignment.

2.2 Sound Source Localization

Sound source localization aims to identify the spatial coordinates of sounding regions in visual scenes by jointly analyzing audio signals and visual cues [Park *et al.*, 2025; Sun *et al.*, 2023; Huang *et al.*, 2023]. Among state-of-the-art methods, ACL not only predicts precise sound source locations but also generates frame-level heatmaps that encode the relative intensity and spatial distribution of sounding regions [Park *et al.*, 2025]. These heatmaps provide rich spatial cues,

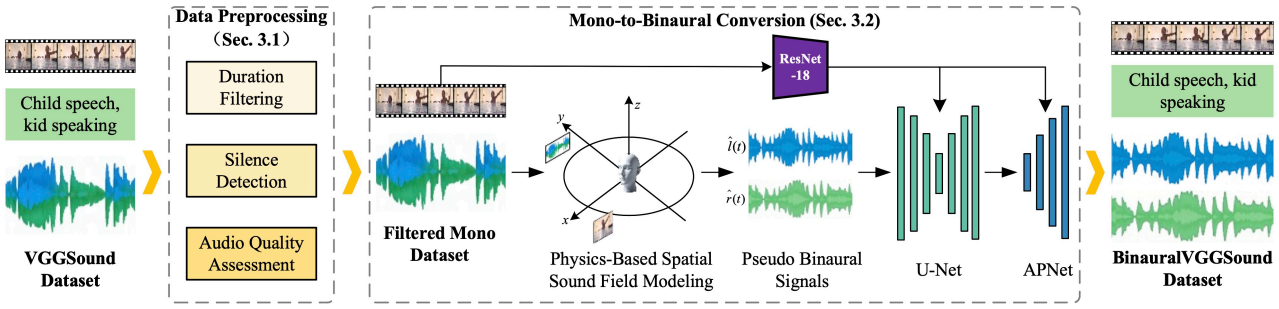

Figure 2: Overall pipeline of BinauralVGGSound dataset construction.

including sound source location and intensity, which can be leveraged to extract spatial information for stereo audio generation.

2.3 Audio Spatialization

Recent advances in visual-guided audio spatialization have provided promising solutions for generating binaural audio [Xu *et al.*, 2021]. Unlike traditional methods that rely solely on acoustic filters or Head-Related Transfer Functions [Zhang *et al.*, 2011], visual-guided approaches [Chen *et al.*, 2025; Xu *et al.*, 2021; Zhou *et al.*, 2020] leverage visual information to infer scene geometry and the relative positions of sound sources. This enables the transformation of monaural audio into binaural audio that better reflects the spatial layout of visual scenes. A representative example is PseudoBinaural [Xu *et al.*, 2021], which synthesizes spatial audio by predicting interaural level and time differences from visual content, effectively transforming monaural audio into plausible stereo or binaural audio.

3 BinauralVGGSound Dataset Construction

Constructing large-scale, visually aligned binaural audio is a critical prerequisite for training spatially aware V2A models. To this end, we introduce **BinauralVGGSound**, a large-scale binaural video–audio dataset for spatial V2A generation. As illustrated in Figure 2, the dataset construction pipeline consists of two main parts: **Data Preprocessing** to ensure temporal alignment and audio quality; and **Mono-to-Binaural Conversion** to facilitate dual-channel spatial V2A modeling.

We adopted VGGSound [Chen *et al.*, 2020] as the data source due to its broad category coverage, reliable audio-visual correspondence, and widespread adoption in prior V2A research [Jeong *et al.*, 2025; Dagli *et al.*, 2025; Xu *et al.*, 2024; Wang *et al.*, 2024; Xing *et al.*, 2024]. The proposed pipeline bypasses the need for costly and hard-to-scale real-world binaural recordings by leveraging an existing mono video–audio dataset to synthesize binaural audio at scale. By doing so, BinauralVGGSound provides audio-visually aligned binaural data for spatial V2A modeling, dramatically lowering data acquisition costs and enabling scalable dataset construction.

3.1 Data Preprocessing

The audio-visual data preprocessing pipeline consists of three key steps: (1) *Duration filtering.* Clips shorter than 10 seconds are discarded to ensure sufficient temporal context for effective model training; (2) *Silence detection.* Videos whose audio contains over 80% silence—identified via an audio activity detector—are removed to preserve meaningful audio–visual correspondence; and (3) *Audio quality assessment.* The Meta Audiobox Aesthetics Toolkit [Tjandra *et al.*, 2025] is employed to filter out degraded or noisy audio samples, resulting in a high-quality dataset appropriate for spatial video-to-audio generation. Detailed preprocessing statistics are provided in the supplementary material.

3.2 Mono-to-Binaural Conversion

Acquiring Ground-Truth (GT) binaural recordings paired with videos is prohibitively expensive and difficult to scale, severely limiting the size of existing datasets. To overcome this, we convert the mono audio from VGGSound into spatially consistent binaural signals using the visual-guided audio spatialization framework PseudoBinaural [Xu *et al.*, 2021].

Specifically, for each mono audio signal $s(t)$, we first apply physics-based spatial sound field modeling to construct pseudo-binaural supervision. The mono signal is represented in the Spherical Harmonic (SH) domain and projected onto a set of virtual sound directions $\boldsymbol{\vartheta}'_m$, corresponding to the m-th spatial component, yielding spatial components $s'_m(t) = \left(D^\top D\right)^{-1} D^\top \boldsymbol{\Psi}(t)$, where D denotes the SH decoding matrix and $\boldsymbol{\Psi}(t)$ the spherical harmonic coefficients.

Pseudo-binaural signals are then obtained via Head-Related Impulse Response (HRIR)-based rendering:

$$\begin{aligned} \hat{l}(t) &= \sum_{m=1}^{M} h_l\left(\boldsymbol{\vartheta}'_m\right) * s'_m(t), \\ \hat{r}(t) &= \sum_{m=1}^{M} h_r\left(\boldsymbol{\vartheta}'_m\right) * s'_m(t), \end{aligned} \tag{1}$$

where M denotes the number of virtual loudspeakers, and $*$ is the convolution operation.

Based on the pseudo-binaural signals $(\hat{l}(t), \hat{r}(t))$, a visual-guided binaural generation model is trained following the PseudoBinaural paradigm. The model employs a U-Net backbone network [Ronneberger *et al.*, 2015] operating in the short-time Fourier transform domain, injects ResNet-18 [He *et al.*, 2016] visual features into the bottleneck layers, and

Dataset	Audio Quality			Audio Spatialization				
	CE ↑	CU ↑	PQ ↑	IACC ↓	ILD ↑	ITD ↑	ISD ↑	IPD ↑
VGGSound [Chen *et al.*, 2020]	3.71	5.39	6.07	-	-	-	-	-
BinauralVGGSound (ours)	**3.74**	**5.44**	**6.09**	**0.93**	**3.08**	**1.70**	**0.07**	**0.32**

Table 2: Analysis of the BinauralVGGSound dataset. Comparison of audio quality and spatialization with the VGGSound dataset.

incorporates APNet-based separation modules [Zhou *et al.*, 2020] to improve sound source discrimination. Once trained, the model is applied to the entire mono VGGSound dataset to generate spatially consistent binaural audio (x^l, x^r). The resulting model-generated binaural audio, together with the original videos and captions, constitutes the final BinauralVGGSound dataset. Its scale and diversity, including the number of clips, total duration, and class coverage, are summarized in Table 1.

3.3 Data Validation

To validate the effectiveness of the constructed BinauralVGGSound dataset, we conduct a comprehensive consistency analysis focusing on *audio quality* and *spatial stereo* characteristics. For *audio quality*, we adopt the Meta Audiobox Aesthetics Toolkit, which assesses three dimensions: Production Quality (PQ), Content Enjoyment (CE), and Content Usefulness (CU). For *spatial stereo*, we compute five widely used binaural metrics: Interaural Cross-Correlation (IACC), Interaural Level Difference (ILD), Interaural Time Difference (ITD), Interaural Spectral Difference (ISD), and Interaural Phase Difference (IPD) [Hernandez-Olivan *et al.*, 2024].

As shown in Table 2, BinauralVGGSound maintains audio quality scores comparable to—and in some cases slightly higher than—the original VGGSound. More importantly, it introduces well-defined spatial cues across all binaural metrics, which are entirely absent in the mono original. Furthermore, Figure 3 illustrates that spatial variations in the binaural audio consistently correspond with visual motion, confirming accurate audio–visual spatial alignment. Additional validation examples are provided in the supplementary material.

These results demonstrate that our construction pipeline successfully enriches the audio with realistic spatial attributes while preserving high perceptual quality, thereby validating the dataset and establishing a solid foundation for spatially aware V2A research.

4 Method

We propose **SpatialV2A**, a dual-branch spatial V2A framework built upon *Conditional Flow Matching (CFM)* and enhanced with visual-guided spatial cues (as shown in Figure 4). The framework first learns a *multimodal joint representation* to align semantics across modalities and provide conditioning for binaural audio generation. Building upon this shared representation, the framework employs a *unimodal dual-channel representation* to decouple the synthesis process into separate left and right branches to model channel-specific spatial variations. A dedicated *visual-guided audio spatialization* module further extracts spatial cues from the video to guide binaural channel differentiation.

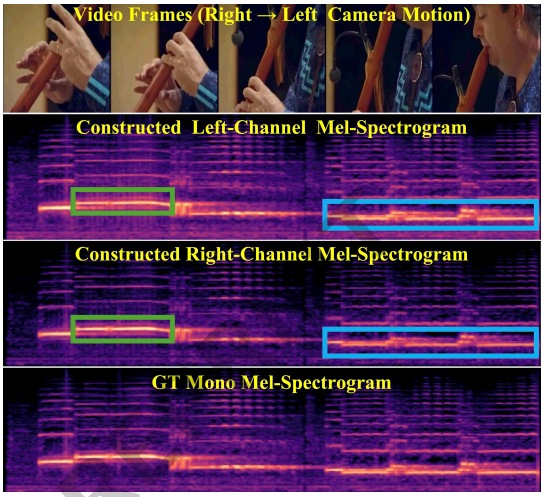

Figure 3: Example visualization from BinauralVGGSound for a flute-playing scene. The transition of bounding boxes from green to blue reflects the estimated sound source movement, demonstrating spatial consistency with the visual motion.

4.1 Preliminaries

CFM is a continuous-time conditional generative framework that learns a condition-dependent velocity field to transport samples from a simple prior distribution to the target data distribution under given conditions [Tong *et al.*, 2024]. Given a target data sample x_1 and a noise sample $x_0 \sim \mathcal{N}(0, I)$, an intermediate state is defined as $x_t = (1-t)x_0 + tx_1, t \in [0, 1]$. The objective of CFM is to learn a conditional velocity field $v_\theta(x_t, t, \mathbf{C})$ parameterized by θ, where $\mathbf{C}$ denotes the conditioning information (*e.g.*, multimodal or unimodal features). This is achieved by minimizing the loss:

$$\mathbb{E}_{x_0, x_1, t, \mathbf{C}} \left\| v_\theta\left(x_t, t, \mathbf{C}\right) - u\left(x_t \mid x_0, x_1\right) \right\|^2, \qquad (2)$$

where $u\left(x_t \mid x_0, x_1\right)$ is the ground-truth conditional velocity.

To generate binaural audio, we extend the standard CFM framework to explicitly model the left and right audio channels. Instead of learning a single velocity field, our framework learns two channel-specific velocity fields, $v_\theta^l(x_t^l, t, \mathbf{C})$ and $v_\theta^r(x_t^r, t, \mathbf{C})$. Given a binaural target audio sample $x_1 = (x_1^l, x_1^r)$ from the constructed BinauralVGGSound dataset and the corresponding noise samples $x_0 = (x_0^l, x_0^r)$, the intermediate states for each channel are constructed via linear interpolation as $x_t^a = (1-t)x_0^a + tx_1^a$, where $a \in \{l, r\}$. Accordingly, the binaural conditional flow matching objective is:

$$\sum_{a \in \{l,r\}} \mathbb{E}_{x_0^a, x_1^a, t, \mathbf{C}} \left\| v_\theta^a\left(x_t^a, t, \mathbf{C}\right) - u\left(x_t^a \mid x_0^a, x_1^a\right) \right\|^2. \qquad (3)$$

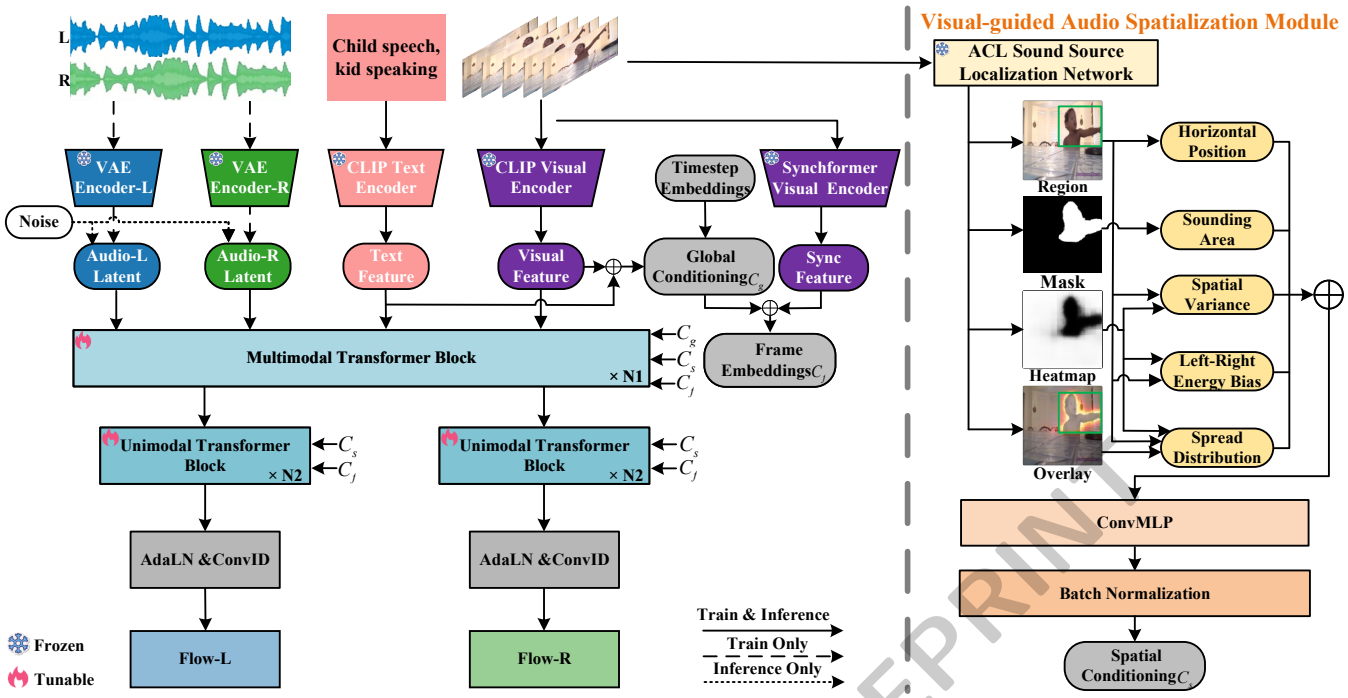

Figure 4: Overall architecture of SpatialV2A. The framework employs a dual-branch design for spatial audio generation (Sec. 4.2) and integrates a visual-guided audio spatialization module (Sec. 4.3) that injects video-derived spatial cues to enhance spatial perception.

4.2 Dual-Branch Spatial Audio Generation

Multimodal Joint Representation

To achieve cross-modal alignment and provide shared conditioning for binaural audio generation, we adopt multimodal joint transformer blocks following the design of MMAudio [Cheng *et al.*, 2025]. We leverage textual F_{text}, visual F_{vis}, and synchronization F_{sync} features for semantic grounding and temporal alignment, implemented via pretrained CLIP [Radford *et al.*, 2021] and Synchformer [Iashin *et al.*, 2024]. The latent representations (x_t^l, x_t^r) are jointly updated by conditioning on six components $\{F_{\text{text}}, F_{\text{vis}}, F_{\text{sync}}, C_s, C_f, C_g\}$. The global conditioning C_g, shared across all transformer blocks, is $C_g = F_{\text{text}} \oplus F_{\text{vis}} \oplus e(t)$, where $e(t)$ denotes the Fourier encoding of the flow timestep. The frame-level conditioning is $C_f = F_{\text{sync}} \oplus C_g$. The spatial conditioning C_s is detailed in Sec. 4.3.

Unimodal Dual-Channel Representation

To model channel-specific spatial characteristics, we introduce unimodal dual-channel blocks that decouple generation into left and right branches after the multimodal stage. Each branch is implemented by a unimodal transformer [Labs *et al.*, 2025] to support independent spatial synthesis while maintaining architectural stability. At each timestep, each branch's latent representation is updated by conditioning on the C_f and C_s.

4.3 Visual-guided Audio Spatialization

This module is designed to provide spatial conditioning C_s for SpatialV2A. Using the pretrained audio-visual localization method ACL [Park *et al.*, 2025], we extract the sounding region, mask, and heatmap from which five key spatial features are computed.

1. Horizontal position S_h represents the relative horizontal placement of the sound source, and is defined as $S_h = \frac{cx}{W}$, where cx denotes the horizontal centroid of the heatmap, and W is the width of the heatmap.

2. Sounding area fraction S_{area} quantifies the proportion of the sounding region within the heatmap and is defined as $S_{\text{area}} = \frac{\sum_{x,y} \text{mask}_{x,y}}{H \times W}$, where x and y denote the pixel coordinates, H denotes the height of the heatmap, and mask is a binary array where pixels with values above a predefined threshold are marked as 1, and 0 otherwise.

3. Heatmap variance S_{var} measures the spatial variance within the heatmap, reflecting the dispersion of the sound source in the spatial domain. It is computed as the sum of the variances in the horizontal $\text{var}_x = \frac{\sum_x (M(x) \cdot (x - cx)^2)}{\sum_x M(x)}$ and vertical $\text{var}_y = \frac{\sum_y (M(y) \cdot (y - cy)^2)}{\sum_y M(y)}$ directions. The overall spatial variance is then given by $S_{\text{var}} = \text{var}_x + \text{var}_y$. Where cy is the vertical centroid of the heatmap, $M(\cdot)$ represents the localization heatmap predicted by ACL.

4. Left-right energy bias S_{LR} captures the energy imbalance between the left and right halves of the visual frame:

$$S_{\text{LR}} = \frac{\sum_{x=W/2+1}^{W} \sum_{y=1}^{H} M(x,y) - \sum_{x=1}^{W/2} \sum_{y=1}^{H} M(x,y)}{\sum_{x=1}^{W} \sum_{y=1}^{H} M(x,y)}. \quad (4)$$

5. Spread distribution shape S_{shape} characterizes the

Method	KL_{PaSST} ↓	KL_{PANNs} ↓	FAD_{PaSST} ↓	FAD_{PaNNs} ↓	IB ↑	DeSync ↓	MOS-Sem ↑	MOS-T ↑
ReWaS	2.47	2.51	693.89	23.46	0.18	1.19	2.31 ± 0.62	2.21 ± 0.44
Frieren	2.62	2.60	490.00	16.96	0.22	0.85	3.40 ± 0.50	2.78 ± 0.58
Seeing&Hearing	2.66	2.78	755.96	28.79	**0.36**	1.20	2.60 ± 0.60	2.53 ± 0.93
MMAudio	<u>1.57</u>	<u>1.59</u>	**303.92**	**8.05**	<u>0.29</u>	<u>0.51</u>	**4.11 ± 0.31**	<u>4.05 ± 0.60</u>
SEE-2-SOUND	3.36	4.39	3807.54	128.03	0.07	1.27	1.60 ± 0.72	1.51 ± 0.66
SpatialV2A (ours)	**1.56**	**1.55**	<u>381.90</u>	<u>12.53</u>	0.28	**0.49**	<u>4.08 ± 0.36</u>	**4.14 ± 0.54**

Table 3: Quantitative and qualitative evaluation across distribution matching, temporal alignment, and semantic alignment. **Bold** indicates the best performance; <u>underlined</u> denotes the second best.

anisotropy of the sound source by measuring its spread along the horizontal and vertical axes. It is defined as $S_{\text{shape}} = \frac{\text{var}_x}{\text{var}_y}$.

These features are concatenated into S_{sound}, projected into a latent space via a Convolutional Multi-Layer Perceptron (ConvMLP) with upsampling, and normalized to form the spatial conditioning:

$$\begin{aligned} S_{\text{sound}} &= S_h \oplus S_{\text{area}} \oplus S_{\text{var}} \oplus S_{\text{LR}} \oplus S_{\text{shape}}, \\ C_s &= \text{LayerNorm}\big(\text{ConvMLP}(\text{UpSample}(S_{\text{sound}}))\big). \end{aligned} \tag{5}$$

The high-dimensional spatial features are incorporated into the flow-matching process of a multimodal transformer, ensuring consistent feature scaling and training stability. This design jointly encodes sound source location, extent, dispersion, energy bias, and directionality, providing essential spatial conditions for binaural audio generation and localization.

5 Experiments

5.1 Experimental Details

Training and Inference. We train our SpatialV2A model on the BinauralVGGSound dataset. During training, audio waveforms are first encoded into latent representations using a pretrained VAE [Kingma and Welling, 2014]. The model is then optimized with the binaural CFM objective (defined in the Preliminaries) to learn the conditional velocity field between noise and target audio latents. During inference, the left and right audio channels are both initialized from Gaussian noise and progressively transformed into binaural audio latents. These predicted latents are decoded into mel-spectrograms via the pretrained VAE and subsequently converted into waveform audio using a BigVGAN vocoder [Lee *et al.*, 2023].

Implementation Details. The model is optimized using AdamW [Kingma and Ba, 2015] with a base learning rate of 1×10^{-4}, a linear warm-up over the first 1K steps, and a total of 300K training iterations. Our architecture employs $N_1 = 6$ multimodal joint transformer blocks and $N_2 = 12$ unimodal dual-channel blocks. Audio is generated at a sampling rate of 16 kHz and represented as 20-dimensional latent codes at 31.25 fps [Jeong *et al.*, 2025; Wang *et al.*, 2024]. All experiments are conducted on three NVIDIA RTX 6000 GPUs with 96 GB of memory.

Baselines. We selected several SOTA methods for V2A as our baselines, including monaural generation models MMAudio [Cheng *et al.*, 2025], ReWaS [Jeong *et al.*, 2025], Frieren [Wang *et al.*, 2024], and Seeing-and-Hearing [Xing *et al.*, 2024], as well as the spatial generation model SEE-2-SOUND [Dagli *et al.*, 2025]. Notably, all baseline models were originally trained on the VGGSound dataset. For a fair evaluation, we evaluate all methods on the same subset of 3,000 videos randomly sampled from the VGGSound test split, which shares identical video IDs with BinauralVGGSound.

5.2 Evaluation Metrics

We evaluate the generated audio across five complementary dimensions.

Distribution Matching quantifies the similarity between the feature distributions of GT and generated audio, reflecting overall fidelity. We compute Fréchet Audio Distance (FAD) and Kullback–Leibler (KL) divergence using PaSST [Koutini *et al.*, 2022] and PANNs [Kong *et al.*, 2020] as feature extractors.

Semantic and Temporal Alignments capture the consistency between audio and visual modalities in terms of semantics and temporal synchronization. We employ ImageBind (IB) [Viertola *et al.*, 2025] and DeSync [Iashin *et al.*, 2024] to estimate semantic and temporal consistency, respectively.

Audio Quality evaluates the perceptual clarity and naturalness of the generated sound using PQ, CE, and CU, together with the Inception Score (IS) [Salimans *et al.*, 2016] computed using PaSST and PANNs.

Spatial Alignment assesses spatial characteristics and localization accuracy in binaural generation. We measure ILD, ITD, ISD, IPD, and IACC [Hernandez-Olivan *et al.*, 2024].

Subjective Evaluation. We conducted a MOS study (1 = poor, 5 = excellent) with twenty invited participants who rated audio samples on high-quality headphones across four dimensions: Spatial Impression (MOS-S), Temporal Alignment (MOS-T), Semantic Alignment (MOS-Sem), and Audio Quality (MOS-AQ).

5.3 Main Results

Table 3 shows the performance of baselines and the proposed method in terms of distribution matching, as well as temporal and semantic alignment. We observe that SpatialV2A achieves the lowest KL divergence and DeSync scores, and ranks first or second on all distribution-related metrics, achieving the highest MOS-T and the second-highest MOS-Sem. It is worth noting that the Seeing&Hearing

Method	IS_{PaSST} ↑	IS_{PaNNs} ↑	CE ↑	CU ↑	PQ ↑	IACC ↓	ILD ↑	ITD ↑	ISD ↑	IPD ↑	MOS-AQ ↑	MOS-S ↑
ReWaS	7.86	8.76	3.54	4.82	5.34	-	-	-	-	-	2.29 ± 0.62	1.26 ± 0.44
Frieren	9.10	8.69	3.65	5.26	5.85	-	-	-	-	-	3.15 ± 0.52	1.29 ± 0.46
Seeing&Hearing	5.39	5.19	3.06	4.49	5.25	-	-	-	-	-	2.42 ± 0.63	1.50 ± 0.54
MMAudio	11.86	12.47	3.88	5.14	5.71	-	-	-	-	-	3.87 ± 0.78	1.78 ± 0.61
SEE-2-SOUND	1.52	1.52	3.03	4.83	5.39	-	-	-	-	-	1.74 ± 0.78	2.49 ± 0.73
SpatialV2A (ours)	**14.36**	**14.91**	**4.09**	**5.77**	**6.07**	**0.51**	**5.69**	**1.79**	**0.49**	**1.12**	**4.21 ± 0.43**	**3.90 ± 0.53**

Table 4: Quantitative and qualitative evaluation of audio quality and spatial alignment.

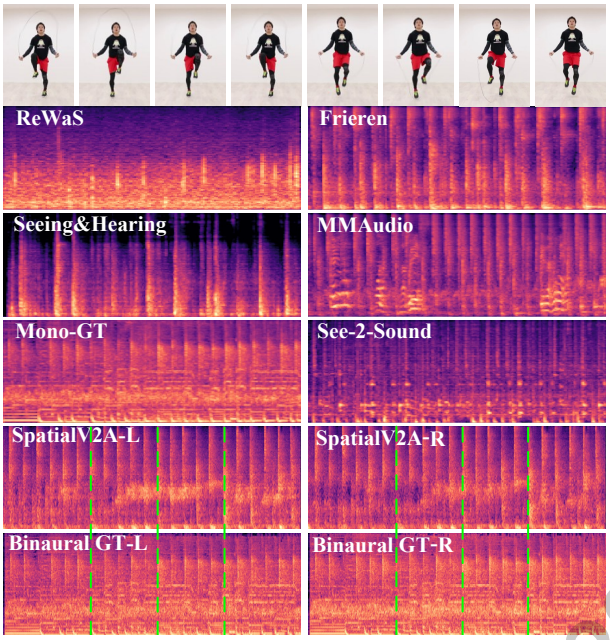

Figure 5: Mel-spectrogram comparison between our method, baselines, original mono GT, and our binaural GT. Our approach better preserves salient audio effects while reducing background noise.

method achieves the highest IB score because it directly optimizes the ImageBind similarity during inference.

Beyond these aspects, Table 4 reports results on audio quality and spatial perception. SpatialV2A achieves the best performance across both objective and subjective metrics. Specifically, it attains the highest scores on audio quality metrics while effectively introducing spatial cues, as evidenced by its performance on spatial metrics. These spatial metrics are not applicable to all baseline methods that generate mono or non-spatial audio. Moreover, SpatialV2A reaches the highest MOS-AQ (4.21) and MOS-S (3.90), significantly outperforming prior works in perceived sound quality and spatial realism. Notably, while monaural audio may convey a sense of immersion or depth through loudness variation or temporal dynamics, it lacks explicit interaural cues required for true spatialization, such as interaural level and time differences.

Figure 5 presents a mel-spectrogram comparison between SpatialV2A and baseline methods. As observed, our approach successfully reconstructs the audio corresponding to the real-world jump-rope scene, exhibiting a more concentrated energy distribution and stable time–frequency structure, which typically indicates higher perceptual clarity and reduced noise interference.

Method	FAD_{PANNs} ↓	IB ↑	DeSync ↓	IS_{PANNs} ↑	IACC ↓	ILD ↑
Ours w/o Spat.	**11.01**	0.26	0.56	13.41	0.53	5.55
SpatialV2A	12.53	**0.28**	**0.49**	**14.91**	**0.51**	**5.69**

Table 5: Ablation study on the proposed visual-guided Audio Spatialization module. We compare our full model (**SpatialV2A**) with a variant without spatialization (**Ours w/o Spat.**).

Overall, these results confirm that SpatialV2A significantly enhances audio quality and spatial perception while maintaining strong temporal and semantic alignment with visual content.

5.4 Ablation Study

To validate the contribution of the proposed *Visual-guided Audio Spatialization* module, we conduct an ablation study by comparing the full SpatialV2A model with a variant that disables spatial conditioning injection (Ours w/o Spat.). As shown in Table 5, removing spatialization slightly improves FAD, indicating marginally improved global distribution matching. In contrast, the full SpatialV2A model achieves higher performance across all critical dimensions, including semantic alignment (IB), temporal synchronization (DeSync), perceptual quality (IS), and binaural spatial metrics (IACC, ILD). These results demonstrate that visual-guided spatialization, despite introducing minor distributional variation, substantially improves temporal coherence, semantic fidelity, and spatial realism.

6 Conclusion

This work addresses the overlooked limitation of spatial perception in V2A generation. We introduce BinauralVGGSound, the first large-scale video-binaural audio dataset, and SpatialV2A, a novel dual-branch framework for visual-guided spatial audio synthesis. Experimental results demonstrate that our approach generates high-fidelity binaural audio with accurate semantic, temporal, and spatial alignment. This significantly enhances the immersive quality and perceptual realism of the synthesized soundscape. For future work, we plan to incorporate captioned datasets such as WavCaps [Mei *et al.*, 2024] and AudioCaps [Kim *et al.*, 2019] to strengthen semantic grounding, and to explore advanced architectures for long-term temporal modeling, enabling coherent binaural synthesis in extended video sequences.

Acknowledgements

This work was supported by the Shandong Postdoctoral Science Foundation (No. SDZZ-ZR-202501291).

References

[Chen *et al.*, 2020] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 721–725, 2020.

[Chen *et al.*, 2025] Yuanhong Chen, Kazuki Shimada, Christian Simon, Yukara Ikemiya, Takashi Shibuya, and Yuki Mitsufuji. Ccstereo: Audio-visual contextual and contrastive learning for binaural audio generation. In *Proceedings of the ACM International Conference on Multimedia*, page 7510–7518, 2025.

[Cheng *et al.*, 2025] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28901–28911, 2025.

[Dagli *et al.*, 2025] Rishit Dagli, Shivesh Prakash, Robert Wu, and Houman Khosravani. See-2-sound: Zero-shot spatial environment-to-spatial sound. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference*, pages 1–2, 2025.

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

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

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Hernandez-Olivan *et al.*, 2024] Carlos Hernandez-Olivan, Marc Delcroix, Tsubasa Ochiai, Naohiro Tawara, Tomohiro Nakatani, and Shoko Araki. Interaural time difference loss for binaural target sound extraction. In *International Workshop on Acoustic Signal Enhancement*, pages 210–214, 2024.

[Huang *et al.*, 2023] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio-visual segmentation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 875–883, 2023.

[Huang *et al.*, 2025] Feizhen Huang, Yu Wu, Yutian Lin, and Bo Du. Spotlighting partially visible cinematic language for video-to-audio generation via self-distillation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1170–1178, 2025.

[Iashin and Rahtu, 2021] Vladimir Iashin and Esa Rahtu. Taming visually guided sound generation. In *British Machine Vision Conference*, 2021.

[Iashin *et al.*, 2024] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5325–5329, 2024.

[Jeong *et al.*, 2025] Yujin Jeong, Yunji Kim, Sanghyuk Chun, and Jiyoung Lee. Read, watch and scream! sound generation from text and video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 17590–17598, 2025.

[Kim *et al.*, 2019] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 119–132, 2019.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

[Kong *et al.*, 2020] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[Koutini *et al.*, 2022] Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Conference of the International Speech Communication Association*, pages 2753–2757, 2022.

[Labs *et al.*, 2025] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

[Lee *et al.*, 2023] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *International Conference on Learning Representations*, 2023.

[Luo *et al.*, 2023] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:48855–48876, 2023.

[Mei *et al.*, 2024] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024.

[Pan *et al.*, 2024] Jiahui Pan, Pengjie Shen, Hui Zhang, and Xueliang Zhang. Innovative directional encoding in speech processing: leveraging spherical harmonics injection for multi-channel speech enhancement. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 6451–6459, 2024.

[Park *et al.*, 2025] Sooyoung Park, Arda Senocak, and Joon Son Chung. Hearing and seeing through clip: A framework for self-supervised sound source localization. *International Journal of Computer Vision*, 2025.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[Salimans *et al.*, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in Neural Information Processing Systems*, 29, 2016.

[Sheffer and Adi, 2023] Roy Sheffer and Yossi Adi. I hear your true colors: Image guided audio generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023.

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

[Tjandra *et al.*, 2025] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

[Tong *et al.*, 2024] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

[Viertola *et al.*, 2025] Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.

[Wang *et al.*, 2024] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Advances in Neural Information Processing Systems*, 37:128118–128138, 2024.

[Xing *et al.*, 2024] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7151–7161, 2024.

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

[Xu *et al.*, 2024] Manjie Xu, Chenxing Li, Xinyi Tu, Yong Ren, Rilin Chen, Yu Gu, Wei Liang, and Dong Yu. Video-to-audio generation with hidden alignment. *arXiv preprint arXiv:2407.07464*, 2024.

[Zhang *et al.*, 2011] Wen Zhang, Mengqiu Zhang, Rodney A Kennedy, and Thushara D Abhayapala. On high-resolution head-related transfer function measurements: An efficient sampling scheme. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):575–584, 2011.

[Zhang *et al.*, 2025] Mengchen Zhang, Qi Chen, Tong Wu, Zihan Liu, and Dahua Lin. Visaudio: End-to-end video-driven binaural spatial audio generation. *arXiv preprint arXiv:2512.03036*, 2025.

[Zhou *et al.*, 2020] Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. Sep-stereo: Visually guided stereophonic audio generation by associating source separation. In *Proceedings of the European Conference on Computer Vision*, pages 52–69, 2020.