V2SA-BENCH: A SOURCE-AWARE BENCHMARK FOR VIDEO-TO-SPATIAL AUDIO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Rapid growth in augmented and virtual reality applications has increased the focus on video-to-spatial audio generation, which conveys immersive cues about objects and their positions in space. However, evaluation efforts remain mono-centric and operate at the global-scene level rather than on individual sources. We introduce V2SA-Bench, a source-aware benchmark for video-to-spatial audio models, built on a human-annotated dataset of 1,274 videos across perspective and 360° videos, focusing on stereo and binaural audio and extending to first-order ambisonics. We define five evaluation dimensions, i.e., semantics, temporal, spatial, acoustics, and perceptual quality, with 12 fine-grained, reference-free, human-aligned metrics. Across 13 baselines, we find that models are strongly center-biased, that spatial alignment is the largest bottleneck, and that performance drops from single to multiple sources. We further find that verifying source presence is essential and that higher perceptual quality does not imply better audio-visual alignment. We hope V2SA-Bench and its interpretable, source-aware metrics offer actionable directions for improving spatial audio generation. We will release our code and benchmark to facilitate future research. The demo webpage is included in the supplementary material.

1 INTRODUCTION

The rapid growth of virtual reality (VR), augmented reality (AR), and metaverse applications, together with teleconferencing and film dubbing, has driven a push toward more immersive spatial audio. Unlike mono audio, spatial audio conveys a rich set of perceptual cues, i.e., directionality, distance, source location, room acoustics, and an overall sense of immersion. Manually authoring such spatial audio for these applications is labor-intensive and requires expert skills, motivating automatic generation directly from visual content. In particular, video-to-spatial audio generation is a natural fit, since the video itself provides strong cues about what sounds are present and where they originate. Building on rapid progress in text- and video-to-audio generation, a growing body of work now aims to synthesize spatial audio, i.e., stereo, binaural, and first-order ambisonics (FOA), directly from video, through both end-to-end models (Tian et al., 2026; Kim et al., 2025; Liu et al., 2025) and multi-stage cascaded pipelines (Gao & Grauman, 2019).

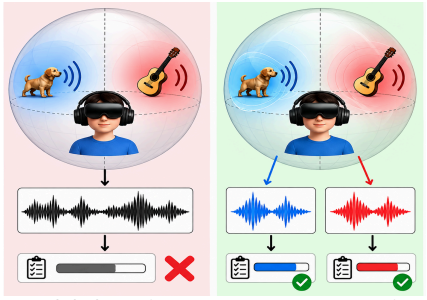

Figure 1: Global metrics overlook source-level mismatches, while our V2SA metrics explicitly evaluate the alignment of each active sound source.

Yet, despite this progress, evaluation benchmarks lag behind. A first limitation is that they have remained overwhelmingly mono audio-centric (Li et al., 2026; Dixit et al., 2026; Zhang et al., 2026). While a few spatial-audio benchmarks exist, they support only a limited set of formats and remain narrow in scope. As a result, existing benchmarks largely assess only what is generated while ignoring where it is localized, cannot verify the multi-channel structure of spatial formats, and fail to reveal spatial errors such as center bias.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison with existing benchmarks for conditional audio generation. Prior benchmarks primarily focus on mono audio generation, whereas V2SA-Bench evaluates video-to-spatial audio generation across 12 dimensions using reference-free metrics. Here, V = video, T = text, A = audio, PoV = perspective video, and FOA = first-order ambisonics.

Benchmark	Input Modality	Output Modality	Bench Size	# Eval Dimensions	Source Aware	All Reference Free metrics	Spatial Metrics	Acoustics Metrics	Audio-visual Alignment
FlatSounds[1]	V (PoV) / T	A (Mono)	185	9	✗	✗	✗	✓	✓
FoleyBench[2]	V (PoV)	A (Mono)	5,000	6	✗	✗	✗	✗	✓
VidAudio-Bench[3]	V (PoV) / T	A (Mono)	1,634	13	✗	✓	✗	✗	✓
TTA-Bench[4]	T	A (Mono)	2,999	7	✗	✓	✗	✗	✗
AudioEval[5]	T	A (Mono)	4,200	5	✗	✓	✗	✗	✗
T2A-EpicBench[6]	T	A (Mono)	100	5	✗	✗	✗	✗	✗
VinTAGe-Bench[7]	V (PoV) / T	A (Mono)	636	9	✗	✗	✗	✗	✓
MMAG[8]	T	A (Mono)	4,000	8	✗	✗	✗	✗	✗
UniHAGen-Bench[9]	V (PoV) / T	A (Mono)	1,003	5	✗	✗	✗	✗	✓
Sphere360-Bench[10]	V (360°)	A (FOA)	3,000	5	✗	✗	✓	✗	✗
V2SA-Bench (Ours)	V (PoV/360°)	A (Stereo/Binaural/FOA)	1,274	12	✓	✓	✓	✓	✓

[1] Li et al. (2026), [2] Dixit et al. (2026), [3] Zhang et al. (2026), [4] Wang et al. (2026), [5] Wang et al. (2025a), [6] Wang et al. (2025c), [7] Kushwaha & Tian (2025), [8] Zheng et al. (2026), [9] Pian et al. (2026), [10] Liu et al. (2025)

A second limitation is that most existing metrics operate at the level of the global scene. While a few recent works have begun to consider source-level cues, each is typically tied to a specific audio format or only a few dimensions. As a result, no benchmark unifies source-aware evaluation across dimensions and spatial formats. Yet humans perceive the world as a mixture of distinct sound sources, such as a fan whirring overhead while footsteps sound from below. This is precisely how professional object-based audio formats such as Dolby Atmos treat sound: as a set of sources, each with its own direction, semantics, and temporal behavior. Evaluation metrics should therefore mirror this perception and compute source-wise measures: given a piano and a guitar in a scene, we should be able to assess the generated audio for each source independently.

A third limitation lies in current evaluation metrics. Common measures such as Fréchet Audio Distance (FAD) (Gui et al., 2024) and KL divergence summarize distribution-level similarity but provide limited diagnostic insight into why a generated audio sample is incorrect or how a model should improve. Other metrics rely on reference audio, which may be unavailable for open-ended generation. We instead focus on objective, reference-free metrics that explicitly measure audio-visual alignment and provide interpretable feedback. Our design follows a simple intuition: *if a human watches a muted video, what should they expect it to sound like?* To make this evaluation well-defined, we curate clips with clearly visible, active sources and unambiguous expected audio.

In summary, current audio evaluation suffers from three shortcomings: (1) it is overwhelmingly mono-centric and overlooks essential spatial cues; (2) it evaluates global scenes rather than individual sources; and (3) it often relies on abstract, distribution-based metrics or reference-dependent measures that offer limited interpretable, source-level feedback to future model developers.

To address these shortcomings, we introduce V2SA-Bench, a benchmark for video-to-spatial audio generation. First, to move beyond mono-centric evaluation, we cover two video formats and three spatial audio formats, focusing on stereo and binaural audio and extending to FOA, over a human-annotated corpus of 1.274 videos (767 perspective and 507 360-degree videos), each manually verified to contain active sound sources and annotated with their corresponding masks. Second, to move beyond global-scene evaluation, and as illustrated in Figure 1, we introduce 12 fine-grained metrics spanning five dimensions: semantic, temporal, spatial, acoustic, and perceptual quality. Most of these metrics are source-aware, operating at the level of individual sound sources, while a scene-level soundscape metric captures the overall acoustic environment and an audio-only metric assesses perceptual quality. Third, to overcome the reliance on distributional or reference-based metrics, our measures are reference-free and human-interpretable. Additionally, to enable a fair and extensive comparison, we evaluate 13 baselines spanning both the end-to-end and multi-stage cascaded paradigms.

Through comprehensive experiments and analyses, we find that current video-to-spatial audio models are strongly center-biased and that spatial alignment remains the largest bottleneck. Multi-stage cascaded pipelines tend to outperform end-to-end models on most dimensions except perceptual quality, though this advantage stems from a high-quality mono first stage rather than the cascaded design itself. Performance also drops consistently from single to multiple sources, and higher perceptual quality does not imply better audio-visual alignment. Finally, we show that verifying source presence is essential and that our metrics align well with human preference.

Our main contributions are summarized as follows:

- **V2SA-Bench**. A comprehensive video-to-spatial audio generation benchmark spanning two video formats and three spatial-audio formats.
- **High-quality dataset**. A manually verified and annotated dataset with active sound sources and their corresponding masks.
- **Source-aware metrics**. Reference-free, human-aligned metrics across five dimensions and 12 fine-grained categories, most of which operate at the level of individual sound sources.
- **Valuable insights**. An extensive evaluation of 13 baselines revealing that current models are strongly center-biased with spatial alignment as the key bottleneck, and that our metrics align well with human judgment.

2 Related works

Spatial audio generation. Early spatial audio was synthesized using signal-processing techniques such as room acoustic simulation (Scheibler et al., 2018), while subsequent deep-learning methods primarily spatialized existing mono audio (Gao & Grauman, 2019). More recently, end-to-end models generate spatial audio directly from video, including video-to-stereo or binaural methods (Tian et al., 2026; Liu et al., 2026a; Karchkhadze et al., 2026; Zhang et al., 2025b; Zhao et al., 2026) and video-to-FOA methods (Kim et al., 2025; Liu et al., 2025; Lei et al., 2026). Cascaded pipelines instead first generate mono audio and then spatialize it using estimated source locations (Gao & Grauman, 2019; Dagli et al., 2025). Our benchmark evaluates such methods whose checkpoints are publicly available. These works also introduce spatial-audio datasets such as FAIR-Play (Gao & Grauman, 2019), Sphere360 (Liu et al., 2025), and YT-Ambigen (Kim et al., 2025); we collect these datasets and manually verify each sample to build the V2SA-Bench dataset.

Evaluation benchmarks. The rapid emergence of generative models has made benchmarks essential for measuring and driving progress. Comprehensive benchmarks exist for image and video generation (Huang et al., 2023; Sun et al., 2025a; Huang et al., 2024), and this maturity has recently extended to mono audio generation(Li et al., 2026; Dixit et al., 2026; Zhang et al., 2026). However, these audio benchmarks operate on the global scene and are mono-centric, and thus cannot assess the spatial and source-wise cues central to spatial audio. Existing spatial metrics are also limited: the DOA-error metric of OmniAudio (Liu et al., 2025) (following ImmerseDiffusion (Heydari et al., 2025)) and the saliency-based metric of ViSAGe (Kim et al., 2025) both require ground-truth audio, while the bin-alignment score of StereoFoley (Karchkhadze et al., 2026), which measures the alignment between object position and the audio energy center, cannot handle multiple simultaneous sources. Despite the growing importance of spatial audio for immersive applications such as virtual environments and world models (Jin et al., 2026), no unified, source-aware benchmark exists. We introduce V2SA-Bench, which evaluates video-to-spatial audio generation models across diverse audio categories with reference-free, human-aligned, and verifiable metrics and compare it with existing benchmarks in Tab. 1.

Source-aware audio generation. Recent work increasingly incorporates object and source-level conditioning into audio generation (Karchkhadze et al., 2026; Li et al., 2025; Guo et al., 2026; Viertola et al., 2026; Yoshida et al., 2026), using visual detection (Guo et al., 2026), segmentation masks (Li et al., 2025; Viertola et al., 2026), or 3D source grounding (Yoshida et al., 2026). Evaluation has similarly begun to consider source-level semantic, temporal, and spatial information (Guo et al., 2026; Viertola et al., 2026; Karchkhadze et al., 2026; Yoshida et al., 2026). V2SA-Bench unifies these aspects through reference-free source-level evaluation of semantic, temporal, spatial, and acoustic alignment across stereo, binaural, and FOA audio. Leveraging recent advances in video segmentation (Carion et al., 2026) and source separation (Shi et al., 2025), we establish cross-modal source correspondences for interpretable evaluation.

3 V2SA-Bench

We introduce V2SA-Bench to understand video-to-spatial audio generation models with a focus on source-aware evaluation. Given a perspective video, we consider two standard spatial audio



Figure 2: Examples and evaluation procedures for each dimension: semantic, temporal, spatial, acoustic, and perceptual quality. Each source-aware metric first verifies source presence, and only then assesses its timing, location, depth, or reverb match.

formats, binaural and stereo. Binaural audio is recorded to simulate human experience while stereo is recorded using left and right channels independently (Wilkins et al., 2023).

3.1 V2SA-Bench Taxonomy and Definitions

Our goal is to evaluate how realistically the generated audio aligns with the input video and its visible sound sources. To this end, we propose explicit audio-visual alignment metrics that quantify semantic, temporal, spatial, and acoustic consistency. Most of these metrics are source-aware, operating at the level of individual sound sources, while the semantic dimension additionally includes a scene-level metric that assesses the overall sound environment. For a more holistic evaluation, we additionally assess the perceptual quality of the generated audio using an audio-only metric.

We briefly define each evaluation dimension and its corresponding evaluation procedure below, with representative examples shown in Fig.2. Most of these procedures share a common count-based structure: for each clip, we first verify source presence and then check whether the corresponding alignment (e.g., temporal or spatial) satisfies a threshold, and we report the proportion of clips that pass as the metric. The soundscape metric follows the same count-based scheme at the scene level, whereas the perceptual-quality metric instead reports an averaged score. The cues used to compute our metrics (e.g., the target sources and their location) are derived from the video and its manual annotations (Sec.3.2) rather than from reference audio. *Please refer to the supplementary for details of definition, full implementation details, prompt templates and examples.*

Semantic alignment. We evaluate semantic correspondence at two levels: *sound events*, measuring whether visually expected sounds are present, and *soundscape*, measuring consistency of the overall sound environment. For sound events, we construct a global audio-visual vocabulary adapted from VGGSound (Chen et al., 2020), which also serves as a source-presence gate for other source-aware

metrics.

• ***Sound events – Single source***. Given a target source (e.g., dog, car, or waves), we rank the vocabulary by PE-AV (Vyas et al., 2026) similarity with the generated audio and check whether the target appears in the top-k ($k = 5$), enabling training-free, zero-shot evaluation.

• ***Sound events – Multiple sources***. For multiple sources (e.g., trumpet and male speech), overlapping or otherwise, we isolate each source using SAM-Audio (Shi et al., 2025) with class prompts and apply the same single-source semantic evaluation.

• ***Soundscape***. We adopt the urban acoustic-scene classes from Wang et al. (2021) (e.g., park, bus, or metro). For evaluation, we use Qwen3-Omni (Xu et al., 2025) for closed-set classification and check whether the ground-truth scene appears among its top-k ($k = 3$) predictions.

Temporal alignment. We evaluate audio-visual synchronization based on the number and temporal characteristics of sound sources. Single-source cases are further categorized into transient events with short, distinct energy peaks and sustained events with continuous activity. For source-wise evaluation, we isolate the target audio using SAM-Audio and mask and crop the video around the corresponding source, reducing interference from irrelevant audio-visual content.

• ***Single source – Sustained***. For sustained sources (e.g., arc welding or violin), we first verify source presence using our single-source semantic metric, then isolate the source with SAM-Audio and use Synchformer to measure synchronization with the corresponding masked visual source over time.

• ***Single source – Transient***. Unlike sustained sounds, transient events provide distinct temporal anchors for synchronization. Similar to sustained case, for a transient event (e.g., gunshots or golf hits), we verify source presence and Synchformer to measure synchronization.

• ***Multiple sources***. We evaluate whether each visual source is temporally aligned with its corresponding generated sound. We first verify source presence using our multi-source semantic metric, then isolate each source with SAM-Audio and compute source-wise Synchformer scores using the corresponding masked and cropped video.

Spatial alignment. For an input video we assess if the generated audio sources in the video are well aligned spatially in horizontal plane.

• ***Single source – Static***. Given a static sound source, this dimension evaluates if the sound source is present in the generated audio and subsequently spatially aligns with the visual position. For example, in Fig. 2, we check for if sitar is present and if it is on the right side. For binaural audio, we estimate azimuth by matching interaural level and phase difference (ILD/IPD) cues with head-related impulse response (HRIR)-derived templates using MagLS (Schörkhuber et al., 2018); for stereo, we infer lateral direction from normalized left–right energy imbalance (Chen et al., 2023).

• ***Single source – Moving***. For a moving sound source, we track the source relative to the camera using SAM3 (Carion et al., 2026) and evaluate whether the estimated audio localization follows its visual trajectory over time. For example, if a wave source moves from left to right in the video, the localized sound should exhibit a corresponding left-to-right trajectory.

• ***Multiple sources***. Given two static sound sources, we evaluate whether their relative visual positions are aligned in the generated audio. We apply SAM-Audio to each channel, average the separated outputs to verify source presence, and, if detected, use their left–right channel energies to estimate relative spatial alignment. For example, a cello appearing left of a guitar should also be localized to its left.

• ***Depth–loudness consistency***. For stable sound sources (e.g., cars, airplanes, or footsteps), audio should become louder as the source approaches the camera and quieter as it moves away. We first verify source presence, then measure whether relative root-mean-square (RMS) loudness is inversely correlated with visual source depth estimated using Depth Pro (Bochkovskiy et al., 2025).

Acoustic consistency. Beyond explicit audio-visual cues, we evaluate whether the generated audio matches the acoustic properties implied by the visual environment (e.g., speech in a church should exhibit high reverberation). We restrict this evaluation to speech, where reliable reverberation estimators are available. After verifying speech presence, we use BERP (Wang et al., 2025b) to estimate the reverberation time (i.e., RT60) and compare it with the visually expected dry or high reverberant condition.

Perceptual quality. For a holistic evaluation of spatial audio generation, we complement audio-visual alignment metrics with the Production Quality (PQ) score from AudioBox Aesthetics (Tjandra et al., 2025) to assess the perceptual quality of the generated audio.

3.2 DATA CURATION

We curate the benchmark through a multi-stage pipeline ensuring annotation quality, dimension relevance, and data diversity. Each dimension requires different filtering process, though roughly following the pipeline shown in Fig. 3. We also ensure that none of the clips come from the train set of training data of baseline models.

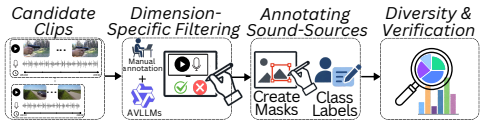

Figure 3: Data curation pipeline. Candidate clips are first extracted from source datasets, followed by dimension-specific filtering, source annotation, and diversity and quality verification.

Candidate clip selection. We curate candidate clips from the source datasets summarized in Tab. 2. For single-source spatial clips, we follow ViSAudio (Zhang et al., 2025b) and crop perspective views from 360° video, decoding their FOA audio to stereo/binaural. This provides controlled source placement in diverse environments. For multi-source spatial clips, since it is difficult to place multiple sources within the cropped field of view, we instead use native stereo/binaural recordings.

Dimension-specific filtering. For sound-event dimensions, we require the sound source to be visually evident and unambiguous; for multi-source cases, the sources must additionally be visually non-overlapping. We filter candidates using an AVLLM, i.e., Qwen3-Omni (Xu et al., 2025), followed by manual inspection. For acoustic-scene and soundscape evaluation, we select clips with clearly identifiable environments (e.g., outdoor spaces or reverberant interiors such as churches), whereas distance clips contain sources that clearly approach or recede from a relatively stationary camera.

Table 2: Source datasets used for candidate clip selection.

Dimension	Source Datasets
Semantics – Single	UnAV-100[1], AVSync48[2]
Semantics – Multi	YouTube-Stereo[3], FAIR-Play[4], MUSIC-Duet[5], AVSBench[6], UnAV-100[1]
Semantics – Scene	TAU AV Scenes[7]
Temporal – Single	UnAV-100[1], AVSync48[2]
Temporal – Multi	YouTube-Stereo[3], FAIR-Play[4], MUSIC-Duet[5], AVSBench[6], UnAV-100[1]
Spatial – Single	Sphere360[8], YT-Ambigen[9]
Spatial – Multi	YouTube-Stereo[3], FAIR-Play[4], MUSIC-Duet[5]
Spatial – Depth	UrbanSAS[10], VGGSound[11]
Acoustics (Reverb)	AVSpeech[12]
Perceptual Quality	UnAV-100[1], YouTube-Stereo[3]

[1] Geng et al. (2023), [2] Zhang et al. (2025a), [3] Sato et al. (2025)
[4] Gao & Grauman (2019), [5] Zhao et al. (2018), [6] Zhou et al. (2022)
[7] Wang et al. (2021), [8] Liu et al. (2025), [9] Kim et al. (2025)
[10] Fuentes et al. (2022), [11] Chen et al. (2020), [12] Ephrat et al. (2018)

Annotations. We adapt a controlled source vocabulary from VGGSound by merging action-specific labels into common source classes (e.g., *dog barking*, *dog howling* → *dog*), yielding 162 classes across humans, animals, vehicles, instruments, household objects, and natural sounds, and manually annotate the active sources in each clip. For source-level spatial and temporal evaluation, we obtain per-source visual masks using SAM3; for classes not well represented by masks (e.g., dart throwing or fireworks), we use bounding boxes instead.

Verification and diversity filtering. Finally, we manually verify semantic correctness, audio-visual consistency, and annotation quality, including source labels and masks, while filtering the set to maintain broad source and scene diversity. This process yields 767 unique perspective videos, with detailed benchmark statistics provided in the appendix.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Baselines. We compare against publicly available video-to-spatial audio generation methods spanning native two-channel and richer multichannel outputs. For native stereo generation, we evaluate AudioX (Tian et al., 2026), ThinkSound (Liu et al., 2026a), PrismAudio (Liu et al., 2026b), and SpatialSonic (Sun et al., 2025b). For binaural generation, we additionally construct a cascaded baseline of two stages: MMAudio (Cheng et al., 2025) for video-to-mono audio generation, followed by Mono2Binaural (M2B) (Gao & Grauman, 2019) for video-guided audio spatialization. We also

Table 3: Comparison of models across our evaluation metrics. Metrics are organized into 5 dimensions (semantics, temporal, spatial, acoustics, and perceptual quality), each further divided into fine-grained sub-categories. GT denotes the ground-truth reference, and the best result in each task (excluding GT) is shown in **bold**. Higher is better for all metrics.

Method	Semantics			Temporal			Spatial				Acoustics	Perceptual Quality
	Single (Events)	Multi. (Events)	Soundscape	Sustained	Transient	Multi	Static (Single)	Moving (Single)	Multi.	Depth		
AudioX[1]	87.9	50.9	**56.0**	32.3	33.9	15.4	16.1	12.1	**16.2**	**40.8**	37.0	**6.5**
ThinkSound[2]	83.3	57.0	37.0	32.3	35.7	33.3	25.0	9.4	15.1	28.6	44.9	6.1
PrismAudio[3]	26.8	26.6	35.0	0.0	5.3	11.2	4.4	0.4	9.3	27.6	30.3	5.5
SpatialSonic[4]	70.3	10.5	49.0	11.2	23.2	8.1	19.1	4.7	4.6	39.4	25.8	5.4
SEE-2-SOUND (Stereo)[5]	16.6	16.4	30.0	2.8	7.1	6.8	4.4	7.0	3.4	1.2	6.7	5.2
SEE-2-SOUND (Binaural)[5]	16.6	16.6	30.0	1.4	0.0	7.1	1.4	6.3	0.0	0.7	7.8	4.6
ViSAGe (Stereo)[6]	29.6	27.4	37.0	2.8	5.3	7.9	17.6	8.8	3.4	15.1	22.4	5.4
ViSAGe (Binaural)[6]	32.4	26.6	37.0	4.2	0.0	8.5	16.1	13.2	10.4	16.5	17.9	5.2
MMAudio[7] + M2B[8]	**98.1**	**70.0**	50.0	**66.2**	**75.0**	**42.0**	**26.4**	**25.3**	15.1	33.8	**48.3**	6.1
GT	99.0	77.7	78.0	70.4	64.2	65.2	77.2	70.2	68.6	72.8	95.5	6.3

[1] Tian et al. (2026), [2] Liu et al. (2026a), [3] Liu et al. (2026b), [4] Sun et al. (2025b), [5] Dagli et al. (2025), [6] Kim et al. (2025), [7] Cheng et al. (2025), [8] Gao & Grauman (2019)

include methods that generate richer spatial formats and decode them to two-channel audio: ViSAGe (Kim et al., 2025), an end-to-end model, generates FOA, whereas SEE-2-SOUND (Dagli et al., 2025), a cascaded pipeline, generates 5.1-channel audio. Both are converted to stereo and binaural for evaluation. Since our metrics are reference-free, we additionally evaluate ground-truth audio. We restrict comparisons to models with publicly available checkpoints, excluding methods such as StereoFoley (Karchkhadze et al., 2026), ViSAudio (Zhang et al., 2025b) and SwanSphere (Lei et al., 2026), whose checkpoints or training pipelines are not publicly released. We inference each of these models only using the video. Any model with input text is kept to empty(""), so that for fair comparison each model only inferences from video input.

4.2 Experimental results

We show our experimental results in Tab. 3. Our key observations are summarized below.

Video-to-mono generation followed by video-guided spatialization shows a strong trend. The cascaded MMAudio + M2B method achieves the best results on 8 of 12 dimensions, ahead of native stereo and richer spatial-format generation models. Notably, being cascaded alone is not sufficient: SEE-2-SOUND, which grounds visual sources in 3D, generates per-source mono audio with CoDi (Tang et al., 2023), and spatializes it via room simulation, is also cascaded yet remains among the lowest-scoring. This suggests that the advantage stems from a high-quality mono first stage whose fidelity the video-guided spatialization preserves while aligning the sound with the video.

Spatial reasoning remains a major bottleneck. The gap between GT and the best baseline is highly uneven across dimensions: it is negligible for single-source semantics (99.0 vs. 98.1) but drastic for spatial alignment (e.g., 77.2 vs. 26.4 for static and 68.6 vs. 16.2 for multi-source). This reveals spatial alignment as the most challenging dimension, showing that current models struggle to localize sound sources according to the visual scene. Moreover, richer spatial-audio formats bring no clear advantage: FOA- and 5.1-based models (ViSAGe, SEE-2-SOUND) perform on par with, and sometimes below, native stereo/binaural models on spatial dimensions.

Increasing the number of sources hurts performance. Moving from single to multiple sources degrades alignment across all dimensions. Taking the overall best model MMAudio + M2B as a representative example, its scores drop from single- to multi-source settings for semantics (98.1 $\rightarrow$ 70.0), temporal (66.2/75.0 for sustained/transient $\rightarrow$ 42.0), and spatial (26.4/25.3 for static/moving $\rightarrow$ 15.1) alignment. The same trend holds for the other models as well. This indicates that strong single-source performance does not extend to multi-source scenes, highlighting the need for source-wise evaluation rather than extrapolating scene-level capability from a collection of individual sources.

Higher perceptual quality does not imply better alignment. AudioX achieves the best PQ (6.5) but only mid-tier alignment scores (e.g., 16.1 on static spatial, 32.3 on sustained temporal), whereas MMAudio + M2B, with a lower PQ (6.1), leads most alignment dimensions (26.4 and 66.2, respectively). Perceptual quality and audio-visual alignment are thus distinct capabilities.

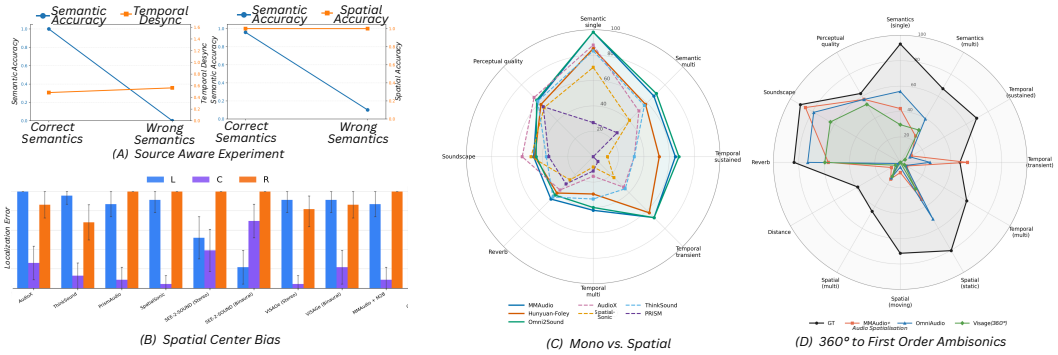

Figure 4: Additional analyses. (A) Under semantic perturbations, temporal and spatial scores stay high while semantic accuracy drops, showing that source-presence checking is essential. (B) Localization error is lowest at the center and grows toward the sides, revealing a spatial center bias. (C) Mono vs. spatial generation across non-spatial metrics. (D) Results on the 360° → FOA setting.

4.3 Additional Analysis

Source presence checking is important. Existing audio-visual temporal and spatial alignment metrics are primarily designed for recorded audio, where the presence of the correct sound source is typically assumed. Directly applying these metrics to generated audio can therefore reward outputs that are temporally or spatially aligned but semantically incorrect. To illustrate this issue, we construct two controlled perturbation settings. For temporal alignment, we use ControlFoley (Yang et al., 2026) to generate audio with incorrect source semantics while preserving similar temporal dynamics (e.g., generating bicycle-acceleration audio for a video of a crying baby). For spatial alignment, we replace a spatially localized sound source with a semantically different source while preserving its spatial position. As shown in Fig. 4 (A), semantic accuracy drops substantially under these perturbations, while temporal and spatial alignment scores remain nearly unchanged. This demonstrates that high temporal or spatial alignment alone does not guarantee that the correct sound source is present, motivating source-presence checking before source-level evaluation.

Spatial center bias. Figure 4 (B) reveals a pronounced center bias across current models: localization error is consistently lowest for centrally located sources and grows toward the left and right. This suggests that models often default toward central spatial predictions rather than faithfully following the visual source direction. A likely contributor is directional imbalance in the training data, highlighting source-direction rebalancing as a concrete future step for improving spatial generation.

Spatial vs. mono audio generation. Our source-aware metrics can be used for mono audio generation except the spatial metrics. On these metrics, we compare native stereo generation and mono audio generation methods in Fig. 4 (C). We observe that mono audio generation models perform better in most dimensions except perceptual quality and soundscape. This supports our earlier observation that the strength of cascaded pipelines stems from a high-quality mono first stage, whose alignment is preserved by the subsequent video-guided spatialization.

Human evaluation. We conducted blinded pairwise comparisons across all 12 evaluation dimensions and three models: AudioX, SpatialSonic, and MMAudio + M2B. In each trial, participants listened to the outputs from two anonymized models and selected which better satisfied the specified evaluation criterion, with additional options for *Both are similar* and *Can't judge*. Model and trial order were randomized, as was dimension order in the combined study. Participants were instructed to use headphones and evaluate only the requested attribute. After data collection, we retained participants who completed all 15 trials for a given dimension and collected at least five completed evaluations per dimension. Overall, the study comprised 5–10 participant sessions per dimension, totaling 79 completed dimension-level sessions. Human preferences were scored as win $= 1$, loss $= 0$, and tie $= 0.5$, while *Can't judge* responses were excluded. As shown in Fig. 5, on our human study, human preferences showed strong agreement with our automatic metrics, with most dimensions achieving correlations above 0.8.

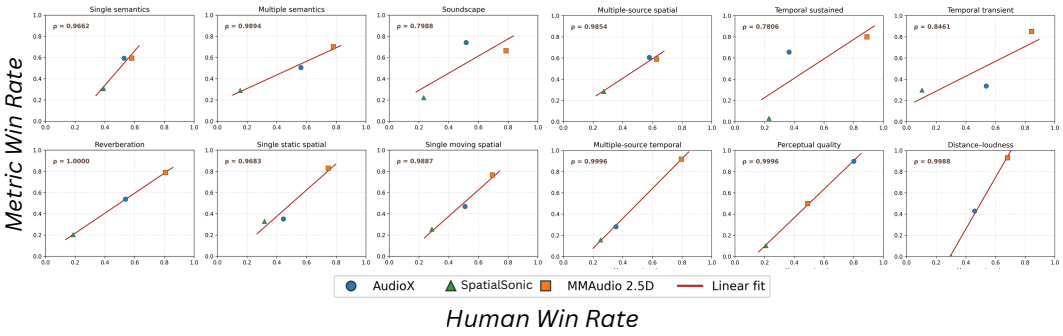

Figure 5: Human study. Per-dimension win rates from our automatic metrics (*Metric Win Rate*) and from blinded human pairwise comparisons (*Human Win Rate*). The strong agreement (most dimensions > 0.8 correlation) indicates that our metrics align well with human preference.

4.4 Extending to 360° videos to FOA

Our source-aware metrics extend naturally beyond perspective videos to the 360° video → FOA task. The metrics remain unchanged, except that spatial localization is performed with a source localization model (Shimada et al., 2025) on the FOA signal, in place of the stereo/binaural estimators used for perspective videos. We source 360° clips from Sphere360 (Liu et al., 2025), YT-Ambigen (Kim et al., 2025), YT360-EyeTracking (Cokelek et al., 2025), and YT-All (Morgado et al., 2018), and apply a curation process similar to the perspective setting, yielding 507 unique 360° videos.

We evaluate two native FOA audio generation models, OmniAudio (Liu et al., 2025) and ViSAGe (Kim et al., 2025). OmniAudio natively operates on 360° video, whereas the original ViSAGe takes a perspective video with a target direction as input. For a fair comparison under our 360° setting, we follow the OmniAudio setup and train ViSAGe directly on 360° video. We additionally consider cascaded baselines that decompose spatial audio generation into video-to-mono generation followed by audio spatialization. For the first stage, the equirectangular 360° video is resized to the input resolution required by the video-to-mono audio generation model (Cheng et al., 2025). For spatialization, we employ SalViT (Cokelek et al., 2025) to estimate the salient source location at 100-ms intervals and spatialize the generated mono audio according to these time-varying locations.

We show the results in Fig. 4 (D). The 360° → FOA setting largely mirrors the perspective setting: the cascaded mono-then-spatialize pipeline again performs strongly, while the end-to-end OmniAudio remains competitive. Spatial reasoning stays the main bottleneck, most notably on moving sources, and the other observations also hold here.

5 Conclusion

We introduced V2SA-Bench, a source-aware benchmark for video-to-spatial audio generation that moves beyond the mono-centric, global-scene, and reference-based evaluation of prior work. It spans two video formats (perspective and 360°) and three spatial-audio formats, focusing on stereo and binaural audio and extending to FOA, over a human-verified corpus of 1,274 annotated clips. We proposed 12 fine-grained, reference-free, human-aligned metrics across five dimensions (semantics, temporal, spatial, acoustics, and perceptual quality), most operating at the level of individual sound sources. Evaluating 13 end-to-end and cascaded baselines, we found that current models are strongly center-biased and that spatial alignment is the largest bottleneck. Cascaded pipelines lead on most alignment dimensions, though this stems from a high-quality mono first stage rather than the cascaded design itself, while end-to-end models yield higher-fidelity audio, showing that perceptual quality and alignment are distinct. Performance also drops markedly from single to multiple sources, exposing multi-source synchronization and localization as a major bottleneck. Finally, verifying source presence proves essential, as source-aware metrics differ substantially from source-agnostic ones and align well with human judgment. We hope V2SA-Bench offers actionable directions for the next generation of video-to-spatial audio models.

AI USE STATEMENT

In this work, we used generative AI tools to assist with paper refinement, baseline code interpretation, visualization script development, and data filtering. We did not use generative AI tools for generating research ideas or citations and references. All AI-assisted content was carefully reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

We followed ethical practices in our human study for evaluating human–metric alignment. Each annotator was automatically assigned a random identifier, and no names or other personally identifiable information were collected. Our benchmark is curated from publicly available YouTube videos; to respect content ownership and platform policies, we will release only the corresponding video URLs and timestamps.

REPRODUCIBILITY STATEMENT

We will release the YouTube URLs, source vocabulary, and annotated masks to support reproducibility and future extension of the benchmark to new methods.

REFERENCES

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, volume 2025, pp. 75602–75637, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Novel-view acoustic synthesis. In *CVPR*, 2023.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911. IEEE, 2025.

Mert Cokelek, Halit Ozsoy, Nevrez Imamoglu, Cagri Ozcinar, Inci Ayhan, Erkut Erdem, and Aykut Erdem. Spherical vision transformers for audio-visual saliency prediction in 360° videos. *IEEE transactions on pattern analysis and machine intelligence*, 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 Posters*, 2025.

Satvik Dixit, Koichi Saito, Zhi Zhong, Yuki Mitsufuji, and Chris Donahue. FoleyBench: A benchmark for video-to-audio models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 14512–14516. IEEE, 2026.

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

Magdalena Fuentes, Bea Steers, Pablo Zinemanas, Martin Rocamora, Luca Bondi, Julia Wilkins, Qianyi Shi, Yao Hou, Samarjit Das, Xavier Serra, et al. Urban sound & sight: Dataset and benchmark for audio-visual urban scene understanding. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 141–145. IEEE, 2022.

Ruohan Gao and Kristen Grauman. 2.5D visual sound. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 324–333. IEEE, 2019.

Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22942–22951. IEEE, 2023.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 1331–1335, 2024.

Wei Guo, Heng Wang, Jianbo Ma, and Weidong Cai. Gotta hear them all: Towards sound source aware audio generation. In *AAAI*, 2026.

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

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. VBench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Derong Jin, Xiyi Chen, Ming C Lin, and Ruohan Gao. Sonoworld: From one image to a 3d audio-visual scene. *arXiv preprint arXiv:2603.28757*, 2026.

Tornike Karchkhadze, Kuan-Lin Chen, Mojtaba Heydari, Robert Henzel, Alessandro Toso, Mehrez Souden, and Joshua Atkins. Stereofoley: Object-aware stereo audio generation from video. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16027–16031. IEEE, 2026.

Jaeyeon Kim, Heeseung Yun, and Gunhee Kim. Visage: Video-to-spatial audio generation. In *International Conference on Learning Representations*, volume 2025, pp. 46404–46424, 2025.

Saksham Singh Kushwaha and Yapeng Tian. Vintage: Joint video and text conditioning for holistic audio generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13529–13539. IEEE, 2025.

Ke Lei, Yu Zhang, Changhao Pan, Xueyi Pu, Wenxiang Guo, Ruiqi Li, and Zhou Zhao. Towards streaming synchronized spatial audio generation via autoregressive diffusion transformer. *arXiv preprint arXiv:2605.30940*, 2026.

Tingle Li, Baihe Huang, Xiaobin Zhuang, Dongya Jia, Jiawei Chen, Yuping Wang, Zhuo Chen, Gopala Anumanchipalli, and Yuxuan Wang. Sounding that object: Interactive object-aware image to audio generation. *arXiv preprint arXiv:2506.04214*, 2025.

Tingle Li, Siddharth Gururani, Kevin J Shih, Gantavya Bhatt, Sang-gil Lee, Zhifeng Kong, Arushi Goel, Gopala Anumanchipalli, and Ming-Yu Liu. Benchmarking single-factor physical video-to-audio generation. *arXiv preprint arXiv:2605.30339*, 2026.

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

Huadai Liu, Kaicheng Luo, Jialei Wang, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. Thinksound: Chain-of-thought reasoning in multimodal llms for audio generation and editing. *Advances in Neural Information Processing Systems*, 38:78526–78552, 2026a.

Huadai Liu, Kaicheng Luo, Wen Wang, Qian Chen, Peiwen Sun, Rongjie Huang, Xiangang Li, Jieping Ye, and Wei Xue. Prismaudio: Decomposed chain-of-thought and multi-dimensional rewards for video-to-audio generation. In *International Conference on Learning Representations*, volume 2026, pp. 25176–25204, 2026b.

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

Weiguo Pian, Saksham Singh Kushwaha, Zhimin Chen, Shijian Deng, Kai Wang, Yunhui Guo, and Yapeng Tian. Omnisonic: Towards universal and holistic audio generation from video and text. *arXiv preprint arXiv:2604.04348*, 2026.

Tomoya Sato, Yusuke Sugano, and Yoichi Sato. Audio-visual localization based on spatial relative sound order. *Machine Vision and Applications*, 36(4):90, 2025.

Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 351–355. IEEE, 2018.

Christian Schörkhuber, Markus Zaunschirm, and Robert Höldrich. Binaural rendering of ambisonic signals via magnitude least squares. In *Proceedings of the DAGA*, volume 44, pp. 339–342, 2018.

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, et al. Sam audio: Segment anything in audio. *arXiv preprint arXiv:2512.18099*, 2025.

Kazuki Shimada, Kengo Uchida, Yuichiro Koyama, Takashi Shibuya, Shusuke Takahashi, Yuki Mitsufuji, and Tatsuya Kawahara. Open-vocabulary sound event localization and detection with joint learning of clap embedding and activity-coupled cartesian doa vector. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2946–2960, 2025.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-compbench: A comprehensive benchmark for compositional text-to-video generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416. IEEE, 2025a.

Peiwen Sun, Sitong Cheng, Xiangtai Li, Zhen Ye, Huadai Liu, Honggang Zhang, Wei Xue, and Yike Guo. Both ears wide open: Towards language-driven spatial audio generation. In *International Conference on Learning Representations*, volume 2025, pp. 42606–42649, 2025b.

Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099, 2023.

Zeyue Tian, Zhaoyang Liu, Yizhu Jin, Ruibin Yuan, Liumeng Xue, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: a unified framework for anything-to-audio generation. In *International Conference on Learning Representations*, volume 2026, pp. 110625–110651, 2026.

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.

Ilpo Viertola, Vladimir Iashin, and Esa Rahtu. Video object segmentation-aware audio generation. *International Journal of Computer Vision*, 134(7):326, 2026.

Apoorv Vyas, Heng-Jui Chang, Cheng-Fu Yang, Po-Yao Huang, Luya Gao, Julius Richter, Sanyuan Chen, Matthew Le, Piotr Dollár, Christoph Feichtenhofer, et al. Pushing the frontier of audio-visual perception with large-scale multimodal correspondence learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30172–30182, 2026.

Hui Wang, Jinghua Zhao, Junyang Cheng, Cheng Liu, Yuhang Jia, Haoqin Sun, Jiaming Zhou, and Yong Qin. Audioeval: Automatic dual-perspective and multi-dimensional evaluation of text-to-audio-generation. *arXiv preprint arXiv:2510.14570*, 2025a.

Hui Wang, Cheng Liu, Junyang Chen, Haoze Liu, Yuhang Jia, Shiwan Zhao, Jiaming Zhou, Haoqin Sun, Hui Bu, and Yong Qin. Tta-bench: A comprehensive benchmark for evaluating text-to-audio models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33512–33520, 2026.

Lijun Wang, Yixian Lu, Ziyan Gao, Kai Li, Jianqiang Huang, Yuntao Kong, and Shogo Okada. Berp: A blind estimator of room parameters for single-channel noisy speech signals. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2215–2231, 2025b.

Shanshan Wang, Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A curated dataset of urban scenes for audio-visual scene analysis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630. IEEE, 2021.

Zehan Wang, Ke Lei, Chen Zhu, Jiawei Huang, Sashuai Zhou, Luping Liu, Xize Cheng, Shengpeng Ji, Zhenhui Ye, Tao Jin, et al. T2a-feedback: Improving basic capabilities of text-to-audio generation via fine-grained ai feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23535–23547, 2025c.

Julia Wilkins, Magdalena Fuentes, Luca Bondi, Shabnam Ghaffarzadegan, Ali Abavisani, and Juan Pablo Bello. Two vs. four-channel sound event localization and detection. *arXiv preprint arXiv:2309.13343*, 2023.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Jianxuan Yang, Xinyue Guo, Zhi Cheng, Kai Wang, Lipan Zhang, Jinjie Hu, Qiang Ji, Yihua Cao, Yihao Meng, Zhaoyue Cui, Mengmei Liu, Meng Meng, and Jian Luan. Controlfoley: Unified and controllable video-to-audio generation with cross-modal conflict handling, 2026. URL `https://arxiv.org/abs/2604.15086`.

Masaki Yoshida, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Scene2sound: Auditory-grounded soundscape generation for 3d gaussian worlds. *arXiv preprint arXiv:2608.00463*, 2026.

Lin Zhang, Zefan Cai, Yufan Zhou, Shentong Mo, Jinhong Lin, Cheng-En Wu, Yibing Wei, Yijing Zhang, Ruiyi Zhang, Wen Xiao, et al. Scaling up audio-synchronized visual animation: An efficient training paradigm. *arXiv preprint arXiv:2508.03955*, 2025a.

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*, 2025b.

Qian Zhang, Yuqin Cao, Yixuan Gao, and Xiongkuo Min. VidAudio-Bench: Benchmarking V2A and VT2A generation across four audio categories. *arXiv preprint arXiv:2604.10542*, 2026.

Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *European conference on computer vision*, pp. 587–604. Springer, 2018.

Lei Zhao, Rujin Chen, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. Foleyspace: Vision-aligned binaural spatial audio generation. *IEEE Transactions on Multimedia*, 2026.

Zihao Zheng, Xuenan Xu, Jiahao Mei, Yixuan Li, Minghao Lv, Wen Wu, Chao Zhang, and Mengyue Wu. Mmag: A multi-control mixed audio generation benchmark. *arXiv preprint arXiv:2608.06900*, 2026.

Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2022.

A APPENDIX

A.1 ADDITIONAL DETAILS ON EVALUATION DIMENSIONS

In this section, we provide additional details on the evaluation dimensions, metric formulations, and the selection of metric hyperparameters.

Semantic Alignment. We evaluate whether the generated audio contains the sound source depicted in the input video. We adopt PE-AV as an open-vocabulary audio–text similarity model. For each generated audio clip, we compute its similarity to textual descriptions of all source classes in our predefined vocabulary. The complete source vocabulary is provided in Table 6. Classes are ranked according to the similarity between their text embeddings and the generated audio embedding, and the top-K predictions are retained for evaluation.

Prompt selection. Since audio–text similarity can be sensitive to prompt formulation, we evaluate several prompt templates on ESC-50. As shown in Table 4, the template "I hear {`class label`} in this audio" achieves the highest Top-1 accuracy, and we therefore use it for all semantic evaluations.

Table 4: Effect of text prompt formulation on PE-AV classification accuracy on ESC-50.

PE-AV Prompt	Top-1 Acc.
{`class label`}	89.75
I can hear {`class label`}	90.75
This is an audio of {`class label`}	91.00
This is {`class label`}	82.50
This is a sound of {`class label`}	90.50
I hear {`class label`} in this audio	**93.75**

For a generated audio clip a_i, let $\mathcal{T}_K(a_i)$ denote its top-K predicted source classes. Let $\mathcal{G}(y_i)$ contain the ground-truth source class y_i together with its predefined equivalent classes. We define the correctness of each clip as

$$c_i = \begin{cases} 1, & \mathcal{T}_K(a_i) \cap \mathcal{G}(y_i) \neq \emptyset, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The semantic alignment score is then computed as the proportion of correctly matched clips:

$$S_{\text{sem}} = \frac{1}{N} \sum_{i=1}^{N} c_i. \tag{2}$$

A higher score indicates that the generated audio more consistently preserves the sound source depicted in the input video.

Selection of K. We determine K using a randomly sampled held-out set of 30 clips. As shown in Table 5, Top-5 retrieval correctly recovers the target semantic class for all validation samples, while increasing K to 10 provides no further improvement. We therefore set $K = 5$ throughout our experiments.

Table 5: Effect of K on semantic matching accuracy on a held-out validation set.

Top-K	Accuracy
1	83.3%
3	96.7%
5	100.0%
10	100.0%

Using Top-5 retrieval also provides robustness to the inherent ambiguity of audio semantics. Acoustically similar classes may receive comparable PE-AV similarity scores even when their visual identities are distinct. For example, sounds such as *pig oinking* and *snoring* may exhibit similar acoustic characteristics. Allowing a small set of highly ranked candidates, together with predefined equivalent classes, avoids overly penalizing such ambiguous cases.

Finally, unlike a fixed closed-set audio classifier, this formulation supports open-vocabulary evaluation: new source classes can be incorporated simply by adding their textual descriptions to the vocabulary, without retraining the classifier. This makes the metric particularly suitable for evaluating diverse, in-the-wild video-to-audio generation.

• ***Sound Events – Multiple Sources***. For videos containing multiple sound sources, we evaluate each ground-truth source independently. For each source, we use its class label as a text prompt to extract the corresponding component from the generated audio using SAM-Audio. We then apply PE-AV to the extracted audio and retrieve its top-5 predicted classes. A source is considered correctly identified if its ground-truth class, or a predefined equivalent (eg. tractor and bus), appears among these predictions.

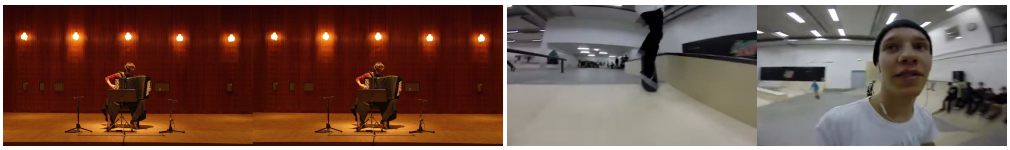

Figure 6: Example of single and multiple source semantics sound events.

For each video, we compute the fraction of correctly identified sources and then average this quantity across all videos:

$$S_{\text{multi}} = \frac{1}{N} \sum_{i=1}^{N} \frac{\#\text{ correctly identified sources in video } i}{M_i}, \tag{3}$$

where M_i denotes the number of ground-truth sources in video i. This video-level averaging ensures that each video contributes equally to the final score, irrespective of the number of sources it contains.

Since SAM-Audio itself uses the source label as a text prompt, one potential concern is that source extraction could introduce a semantic bias. However, successful extraction alone is insufficient to obtain a correct score: the resulting audio must independently match the target class under PE-AV ranking over our large source vocabulary. Therefore, spurious agreement between the extraction prompt and the subsequent semantic prediction is unlikely.

For SAM-Audio, we use FP16 precision for source extraction and FP32 precision for PE-AV re-ranking, with four re-ranking candidates. In our ablations, using FP32 for source extraction or increasing the number of re-ranking candidates from four to eight did not provide measurable improvements, so we use the more computationally efficient configuration throughout.

A qualitative example of the multi-source semantic evaluation is shown in Fig. 6.

Soundscape. We evaluate whether the generated audio is consistent with the acoustic environment depicted in the input video using an audio-only scene classifier based on Qwen3-Omni-30B-A3B-Thinking. Our evaluation set contains 100 clips from the TAU Urban Acoustic Scenes dataset, balanced across 10 classes with 10 clips per class: *airport*, *bus*, *metro*, *metro station*, *park*, *public square*, *shopping mall*, *pedestrian street*, *street traffic*, and *tram*.

For each generated soundtrack, Qwen3-Omni analyzes acoustic cues such as enclosure, listener motion, road or rail activity, crowd presence, natural ambience, traffic, and announcements, and ranks the five most likely scene classes. We report Top-3 accuracy, counting a sample as correct if its ground-truth scene appears among the three highest-ranked predictions. Below is the prompt used.

We use Top-3 rather than Top-1 because several urban soundscapes are inherently ambiguous from short audio alone. For example, *metro*, *metro station*, and *tram* may share rail and announcement cues, while *airport*, *shopping mall*, and *metro station* can exhibit similar crowd and reverberation characteristics. Likewise, *public square*, *pedestrian street*, and *street traffic* may contain overlapping traffic and crowd sounds. Requiring exact Top-1 classification can therefore penalize acoustically plausible generations due to evaluator uncertainty rather than a genuine soundscape mismatch. Top-3 provides limited tolerance to this ambiguity while remaining sufficiently selective.

```
Analyze this recording as a closed-set acoustic scene.

First determine only these acoustic attributes. Use "uncertain" whenever direct
evidence is absent:

- enclosure: indoor, outdoor, or uncertain
- listener_motion: moving, stationary, or uncertain
- road_vehicle_evidence: present, absent, or uncertain
- rail_evidence: present, absent, or uncertain
- crowd_activity: low, medium, high, or uncertain
```

Table 6: Sound Source vocabulary used in our benchmark for semantic presence checking.

Transportation & Vehicles. motorcycle, motorboat, airplane, helicopter, train, tractor, vehicle horn, snowmobile, car, car doors, car windows, car wiper, bicycle, bus.

Alarms, Bells & Signals. police siren, ambulance siren, fire truck siren, alarm clock, smoke detector, air horn, bell ringing, wind chime.

Animals, Birds & Insects. bird, goose, dog, cat, frog, lion, chicken, cattle, sheep, pig, elephant, goat, whale, cow, zebra, cricket chirping, insect buzzing, gibbon howling, horse clip-clop.

Human Vocalizations. scream, booing, crowd, whistling, kid speaking, man speaking, woman speaking, baby crying, laughing, coughing, singing choir, child singing, female singing, male singing, baby babbling, beatboxing, burping, nose blowing, slurping, sneezing, sobbing.

Human Actions. clapping, eating, slapping, finger snapping, tap dancing, keyboard typing, mouse clicking, writing sound, footsteps, rope skipping.

Music, Instruments & Ensembles. singing bowl, guitar, bass guitar, double bass, cello, violin, piano, drums, trumpet, trombone, saxophone, flute, clarinet, french horn, cornet, harp, accordion, banjo, mandolin, ukulele, sitar, erhu, theremin, organ, vibraphone, xylophone, glockenspiel, harpsichord, harmonica, didgeridoo, bagpipes, percussion instrument, synthesizer, tambourine, gong, oboe, steelpan, tuning fork, handpan, ghatam, orchestra, bassoon, tabla, cajon, zither.

Sports & Games. basketball bounce, tennis, badminton, volleyball, skiing, skateboarding, pool striking, football kicking, darts, bowling sounds, golf swing.

Household Appliances. vacuum cleaner, electric blender, toilet flushing, electric shaver, washing machine, electric fan, air conditioning noise, hair dryer.

Machines & Tools. chainsaw, printer, lawn mowing, lathe spinning, hammer sound, chopping sound, arc welding.

Household Object Interactions. liquid pouring, opening drawers, closing drawers, scraping, door sound, object dragging, knife sharpening.

Nature, Weather & Water. volcano explosion, water flowing, rain, underwater bubbling, thunder, sea waves, wind blowing, waterfall.

Impacts, Weapons & Fire. fireworks, missile, fire crackling, gun.

```
- nature_activity: low, medium, high, or uncertain
- traffic_activity: low, medium, high, or uncertain
- announcement_evidence: present, absent, or uncertain

Important rules:
- Do not infer a source or location from timbre alone.
- Generic rumble, hum, reverberation, or metallic noise is not rail evidence.
- Rail evidence requires a consistent combination of unmistakable track or
  wheel behavior and motion or station context.
- Do not infer motion unless the audio changes consistently over time.
- Do not invent cues that are not clearly audible.
- Treat all classes as equally likely before listening.

After extracting the attributes, evaluate every candidate independently.
A compatible but non-specific cue provides weak evidence. Contradictory
attributes should strongly lower a candidate.

Rank five distinct labels:
[airport, bus, metro, metro_station, park, public_square, shopping_mall,
 street_pedestrian, street_traffic, tram]

Return only:
{"class_pred":["rank1","rank2","rank3","rank4","rank5"],
 "reason":"Explain rank 1 using only the extracted attributes and explicitly state
 why the closest alternative ranked lower."}
```

Temporal Alignment. We evaluate whether the generated audio contains the correct sound source and whether that source is synchronized with its visual activity. A sample receives temporal credit only when both conditions are satisfied.

We evaluate three settings: *sustained*, *transient*, and *multi-source*. Sustained events remain active over longer periods, whereas transient events correspond to short, localized sound events. Multi-source videos are evaluated independently for each annotated source.

Semantic verification. We first verify whether the target sound is present using PE-AV. The generated audio is compared against our source vocabulary using the prompt "I hear $\{\texttt{class}\}$ in this audio." A source is considered present if its ground-truth class, or an accepted equivalent, appears among the top-5 predictions.

Temporal synchronization. For each target source, we use SAM-Audio to extract the corresponding sound from the generated audio. The visual input is masked to retain only the target source, reducing interference from unrelated motion. We then use DeSync to estimate the audio–visual temporal offset. A source is considered synchronized when

$$d_i \leq 0.7 \text{ s}, \tag{4}$$

where d_i denotes the predicted absolute audio–visual offset. Lower DeSync values indicate better temporal alignment. *Single-source evaluation.* For single-source videos, let s_i indicate whether the target sound is semantically present and let t_i indicate whether it satisfies the temporal threshold. We report the joint semantic–temporal score

$$S_{\text{temp}} = \frac{1}{N} \sum_{i=1}^{N} s_i t_i. \tag{5}$$

Thus, a sample receives credit only when the correct sound is generated and is temporally synchronized with the corresponding visual event. Sustained and transient subsets use the same metric and are reported separately.

Multi-source evaluation. For videos containing multiple sources, we first extract each annotated source independently using SAM-Audio. PE-AV then verifies the semantics of each extracted source, and DeSync measures its synchronization with the corresponding masked visual source. For video i with M_i annotated sources, we compute

$$S_{\text{temp}}^{(i)} = \frac{1}{M_i} \sum_{j=1}^{M_i} s_{ij} t_{ij}, \tag{6}$$

and average this score across videos:

$$S_{\text{temp}}^{\text{multi}} = \frac{1}{N} \sum_{i=1}^{N} S_{\text{temp}}^{(i)}. \tag{7}$$

This source-wise formulation ensures that the correct sound must be synchronized with the correct visual source, while the per-video averaging prevents videos with more annotated sources from receiving greater weight.

Spatial Alignment – Single Source. We evaluate whether the generated sound is spatialized at the same *left*, *center*, or *right* location as the visible source. For static clips, we use the annotated source location directly. For moving clips, we divide the video into one-second windows and compute the median horizontal position of the source mask within each window. The normalized horizontal position is quantized into left, center, and right regions using boundaries of 0.45 and 0.55.

Audio localization. For stereo audio, we estimate the source direction using the normalized left–right energy difference

$$c = \frac{E_L - E_R}{E_L + E_R}, \qquad E_L = \sum_t L_t^2, \quad E_R = \sum_t R_t^2. \tag{8}$$

We compare the observed cue c with the expected cues produced by our stereo renderer for candidate azimuths between -90° and 90°, sampled at 2° intervals, and select the closest matching direction. For binaural audio, we instead estimate direction by matching interaural level and phase differences to HRIR-based directional templates.

The estimated azimuth $\hat{\theta}$ is projected onto the horizontal video coordinate using the camera field of view α:

$$\hat{x} = \frac{1}{2} - \frac{\tan \hat{\theta}}{2 \tan(\alpha/2)}. \tag{9}$$

We then apply the same $0.45/0.55$ boundaries to obtain the predicted left/center/right audio location.

Source-aware spatial score. For clip i, let r_i denote the fraction of evaluated windows for which the predicted audio location agrees with the visual source location. For static clips, r_i corresponds to a single left/center/right comparison; for moving clips, it is averaged over all valid one-second windows.

We additionally use PE-AV to verify that the generated audio contains the target source. Let $s_i = 1$ if the ground-truth class, or an accepted equivalent, appears in the PE-AV top-5 predictions, and $s_i = 0$ otherwise. We report

$$S_{\mathrm{loc}} = \frac{\sum_i s_i r_i}{\sum_i s_i}, \qquad S_{\mathrm{SemLoc}} = \frac{1}{N}\sum_{i=1}^{N} s_i r_i. \tag{10}$$

Here, S_{loc} measures localization accuracy only among clips in which the correct sound source is present, while S_{SemLoc} is the stricter source-aware metric: a clip receives credit only when the correct sound is generated *and* spatialized at the correct visual location.

Spatial Alignment – Multiple Sources. For videos containing multiple sound sources, we evaluate whether the generated audio preserves the *relative left–right ordering* of the visible sources. Each ground-truth source class is first used as a text prompt to extract the corresponding component from the generated two-channel audio using SAM-Audio. We then verify the semantics of each extracted source with PE-AV. Source j receives a semantic hit, $s_j = 1$, if its ground-truth class, or an accepted equivalent, appears among the top-5 predictions.

Visual and audio positions. For each source, the visual horizontal position x_j^V is obtained by computing the median horizontal position of its mask in each frame and then taking the median across the full clip. To estimate its audio position, we use interaural cues from the extracted stereo signal: low-frequency interaural time differences (ITD; 50–620 Hz) and high-frequency interaural level differences (ILD; 1.7–4.6 kHz). These cues are combined to obtain a continuous horizontal audio position x_j^A.

Unlike the single-source metric, which uses left/center/right regions, the multi-source metric directly compares these continuous source positions over the entire clip. For each pair of sources (a, b), we define the relative spatial correctness as

$$r_{ab} = \mathbf{1}\left[(x_b^V - x_a^V)(x_b^A - x_a^A) > 0\right]. \tag{11}$$

Thus, a pair is correct when the source that appears farther to the right in the video is also localized farther to the right in the generated audio.

Both sources must also be semantically identified for the pair to receive source-aware credit. We therefore define

$$s_{ab} = s_a s_b. \tag{12}$$

Across all P source pairs, we report

$$S_{\mathrm{rel}} = \frac{1}{P}\sum_{(a,b)} r_{ab}, \tag{13}$$

$$S_{\mathrm{cond}} = \frac{\sum_{(a,b)} s_{ab} r_{ab}}{\sum_{(a,b)} s_{ab}}, \qquad S_{\mathrm{joint}} = \frac{1}{P}\sum_{(a,b)} s_{ab} r_{ab}. \tag{14}$$

Here, S_{rel} measures relative spatial ordering alone, S_{cond} measures ordering among semantically valid source pairs, and S_{joint} is the stricter source-aware metric, requiring both sources to be correctly identified and correctly ordered in space.

Depth-loudness consistency. We evaluate whether the generated sound becomes louder as the visible source approaches the camera and quieter as it moves away. For each video, Depth Pro estimates the source depth in every frame. We compute the median depth within the annotated source mask and aggregate these estimates into one-second values $d_{i,t}$. In parallel, SAM-Audio extracts the target sound from the generated audio, and its RMS amplitude within each one-second window gives $e_{i,t}$.

A transition is considered correct when depth and audio intensity change in opposite directions:

$$h_{i,t} = \mathbf{1}[(d_{i,t+1} - d_{i,t})(e_{i,t+1} - e_{i,t}) < 0]\,. \tag{15}$$

Thus, a source moving closer while becoming louder, or moving farther while becoming quieter, receives credit. Transitions with no change, or without a valid depth estimate, are not counted as correct.

We additionally verify that the generated clip contains the target sound using PE-AV. Let $s_i = 1$ if the ground-truth class, or an accepted equivalent, appears among the top-10 predictions, and $s_i = 0$ otherwise. Let T_i denote the number of valid transitions in video i.

For semantically correct clips, we first compute the mean distance-consistency score

$$S_{\text{dist}} = \frac{\sum_{i:s_i=1} \left(\frac{1}{T_i} \sum_t h_{i,t} \right)}{\sum_i s_i}. \tag{16}$$

We also report a joint semantic–distance score:

$$S_{\text{SemDist}} = \frac{\sum_i s_i \sum_t h_{i,t}}{\sum_i T_i}. \tag{17}$$

Here, S_{dist} measures depth–loudness consistency only among clips where the target sound is semantically present, giving each such video equal weight. In contrast, S_{SemDist} gives credit only when both the correct sound is present and the corresponding depth–loudness transition is consistent; since it is normalized over all valid transitions, longer clips may contribute more.

Acoustics. We evaluate whether the generated audio exhibits a reverberation level that is consistent with the visible environment. We use BERP to estimate the reverberation time, $\hat{T}_{60}$, i.e., the time required for the sound level to decay by 60 dB. The estimated value is mapped to one of three categories:

$$\hat{c}_i = \begin{cases} \text{dry}, & \hat{T}_{60,i} < 0.3\,\text{s}, \\ \text{high reverb}, & \hat{T}_{60,i} > 0.5\,\text{s}, \\ \text{uncertain}, & \text{otherwise}. \end{cases} \tag{18}$$

A clip passes the reverberation check when the predicted category matches its annotated acoustic condition. Predictions in the intermediate 0.3–0.5 s range are treated as uncertain and receive no credit for either the dry or high-reverberation class.

We additionally verify that the generated audio contains the intended sound using PE-AV. Let $r_i = 1$ if the predicted reverberation class is correct, and let $s_i = 1$ if the ground-truth sound class, or an accepted equivalent, appears among the top-5 PE-AV predictions. Over N clips, we report

$$S_{\text{rev}} = \frac{1}{N} \sum_{i=1}^{N} r_i, \qquad S_{\text{sem}} = \frac{1}{N} \sum_{i=1}^{N} s_i, \qquad S_{\text{SemRev}} = \frac{1}{N} \sum_{i=1}^{N} r_i s_i. \tag{19}$$

Here, S_{rev} measures reverberation consistency alone, while S_{SemRev} is the source-aware joint score, giving credit only when the generated audio contains the intended sound and exhibits the expected reverberation level.

A.2 Benchmark statistics

Figure 7 (A) shows that the benchmark provides broad coverage across semantic, temporal, spatial, acoustic, and perceptual evaluation dimensions, with approximately 50–150 videos per dimension. As some videos are annotated for and reused across multiple dimensions, these per-dimension counts overlap, yielding 767 unique perspective videos overall. Fig. 7 (B) shows that, over the 767 unique videos, the benchmark spans a broad source taxonomy, including music, human sounds, vehicles, impulsive events, sports, tools, alarms, household sounds, and environmental sources. Music constitutes the largest source category, followed by human and vehicle-related sounds. Fig. 7 (C) shows that a substantial subset contains multiple simultaneously active sources, typically two to three per video, enabling explicit source-aware evaluation.

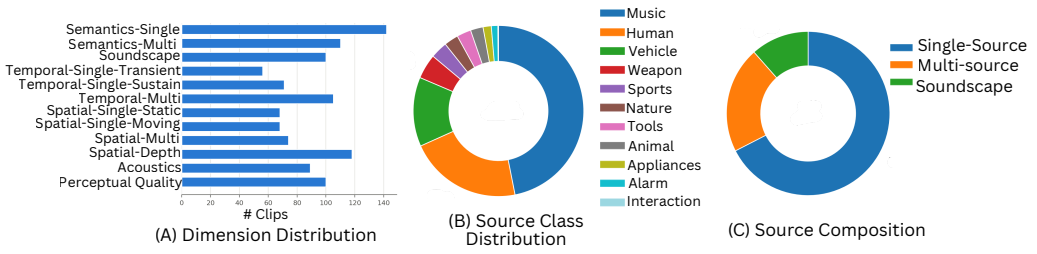

Figure 7: Benchmark stats. (A) Number of clips per evaluation dimension, (B) distribution over source classes, and (C) composition of single-source, multi-source, and soundscape clips.

Table 7: Training data and model scales of the evaluated baseline methods, grouped by task. Since SEE-2-SOUND (Dagli et al., 2025) uses CoDi (Tang et al., 2023) in a zero-shot manner, we report CoDi instead of SEE-2-SOUND. Note that CoDi and AudioX are also trained on tasks other than audio generation; here we report only the training data relevant to audio generation. N/R denotes not reported.

Method	**# of Train Samples**	**Train Data Hours**	**Trainable Params**
Mono Audio Generation			
CoDi (Tang et al., 2023)	4.48M	N/R	N/R
MMAudio (Cheng et al., 2025)	1.13M	2.61k	1.03B
Stereo Audio Generation			
AudioX (Tian et al., 2026)	1.69M	4.99k	1.1B
ThinkSound (Liu et al., 2026a)	N/R	2.53k	1.3B
PrismAudio (Liu et al., 2026b)	N/R	8.13k	518M
SpatialSonic (Sun et al., 2025b)	1.02M	2.85k	N/R
FOA Audio Generation			
ViSAGe (Kim et al., 2025)	0.28M	0.67k	360M
Audio Spatialization			
Mono2Binaural (Gao & Grauman, 2019)	1.50K	4.16	N/R