Cinematic Audio Source Separation Using Visual Cues

Anonymous CVPR submission

Abstract

*Cinematic Audio Source Separation (CASS) aims to decompose mixed film audio into speech, music, and sound effects, enabling applications like dubbing and remastering. Existing CASS approaches are audio-only, overlooking the inherent audio-visual nature of films, where sounds often align with visual cues. We present the first framework for audio-visual CASS (AV-CASS), leveraging visual context to enhance separation quality. Our method formulates CASS as a conditional generative modeling problem using conditional flow matching, enabling multimodal audio source separation. To address the lack of cinematic datasets with isolated sound tracks, we introduce a training data synthesis pipeline that pairs in-the-wild audio and video streams (*e.g.*, facial videos for speech, scene videos for effects) and design a dedicated visual encoder for this dual-stream setup. Trained entirely on synthetic data, our model generalizes effectively to real-world cinematic content and achieves strong performance on synthetic, real-world, and audio-only CASS benchmarks.*

1. Introduction

Cinematic audio is composed of layered sound elements such as speech, music, and sound effects, which collectively enrich storytelling and immersion. The goal of Cinematic Audio Source Separation (CASS) is to separate a mixed movie audio into these three distinct tracks (Fig. 1). CASS enables a wide range of applications, including multilingual dubbing, film remastering, audio editing, and accessibility enhancement. As video streaming platforms continue to grow, the need for CASS becomes increasingly important to enable automatic tools that precisely control individual sound components.

While related audio separation tasks, such as speech separation [8, 24, 62] and music demixing [14, 15, 44, 52], have seen significant progress, CASS remains underexplored. The introduction of the Divide and Remaster (DnR) dataset [48] formalized CASS as a three-way separation problem and initiated new research [57, 59, 60]. However, existing methods are purely audio-based and overlook a defining characteristic of film: its inherently audio-visual nature.

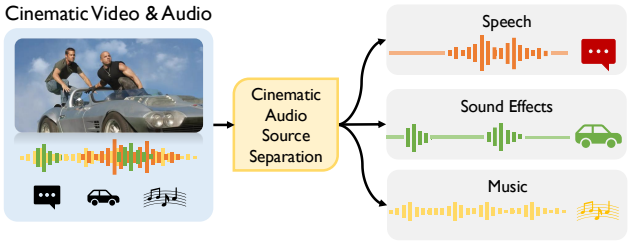

Figure 1. **Illustration of the Cinematic Audio Source Separation (CASS) task.** The audio stream from a movie is separated into distinct tracks: speech, sound effects, and music.

In cinematic content, sound is often tightly coupled with visual events. Speech typically co-occurs with lip movements, and sound effects frequently align with object interactions or visual actions. Prior work in audio-visual learning has shown that using visual context, such as facial motion, body movement, and scene composition, significantly improves audio separation and enhancement [1, 7, 16, 20, 21, 55]. Yet, to our knowledge, visual information has not been utilized in the context of CASS.

The primary motivation of this paper is to achieve cinematic audio source separation using visual cues. Accordingly, we first formulate CASS as a conditional generative modeling problem. We adopt conditional flow matching [39] as our generative framework, which has demonstrated strong performance in both image [17, 27] and audio [22, 31, 35, 58] generation tasks. Our model generates clean speech, sound effects, and music tracks from the mixture, conditioned on both audio and visual inputs.

Second, despite the strong potential of visual cues, applying them to CASS remains challenging. A key obstacle is the lack of publicly available audio-visual datasets with clean source tracks, which are difficult – if not impossible – to obtain from real films. This raises an important question: *Can we leverage individually available in-the-wild audio-visual data to train an effective audio-visual CASS model?* To address this, we propose a novel training data synthesis

strategy. In the absence of paired datasets such as films and their clean sound source tracks, we construct a pipeline that synthetically pairs individually available audio and video sources, *e.g.*, using facial videos for speech and scene video clips for sound effects. This results in a dual-stream video setup that provides controllable, source-specific visual supervision, for which we also design a visual feature extractor to leverage this setup. Importantly, although our model is trained using this synthetic dual-video configuration, we demonstrate that it generalizes effectively to real-world cinematic content without architectural changes, highlighting the practicality and robustness of our training approach.

In summary, our contributions are:

- We introduce the first framework for audio-visual cinematic audio source separation (AV-CASS), incorporating visual cues to separate speech, sound effects, and music in film audio.
- We formulate CASS as a generative task using conditional flow matching, enabling flexible and principled modeling of multimodal audio decomposition.
- We propose a training data synthesis strategy, along with a dedicated visual encoder, to enable training without the need for original source-separated film data.
- We achieve strong performance on synthetic data, real-world cinematic examples, and standard audio-only CASS benchmarks.

2. Related Work

2.1. Cinematic Audio Source Separation

Cinematic Audio Source Separation (CASS) was formalized by [48, 49], introducing the Divide and Remaster (DnR) dataset for the separation of speech, sound effects, and music in film audio. BandIt [59] applied Band-split RNNs [42] to this task, achieving improved performance over earlier methods. More recently, DnRv3 [60] expanded the dataset with multilingual content and introduced mixing strategies aligned with industrial audio production pipelines. DnR-nonverbal [23] further expands the dataset by incorporating nonverbal sounds, such as laughter and screams. While these efforts have advanced audio-based CASS, they remain limited to audio-only learning. To our knowledge, no prior work has explored audio-visual approaches for CASS, despite the inherently multimodal nature of cinematic content. This gap largely stems from the difficulty of acquiring datasets with both isolated audio tracks and temporally aligned video. In contrast, we propose the first audio-visual CASS framework and introduce a training data synthesis pipeline that constructs audio-visual training samples from in-the-wild video sources, without requiring ground-truth source audio with film datasets. Our approach enables effective audio-visual learning and generalizes well to real-world cinematic content.

2.2. Audio-Visual Source Separation

Incorporating visual information has proven highly effective in sound source separation tasks. Prior work has shown that visual cues such as lip movements align strongly with spoken content [10, 43], while facial features provide cross-modal biometric information that enhances speech separation [1, 3, 12, 16, 46, 47]. Beyond speech, visual information from instrument motion or class-level appearance cues has also been leveraged to improve music separation [5, 18, 19, 66]. More generally, recent works have demonstrated that visual context benefits generic sound separation across a wide range of categories [7, 9, 21, 28, 29, 54–56, 63]. Despite these advances, applying audio-visual learning to the complex, multi-source nature of cinematic audio remains largely unexplored. In this work, we introduce audio-visual learning to Cinematic Audio Source Separation (CASS) for the first time. Unlike prior AVSS approaches that condition on a single visual cue (e.g., facial motion [10, 43] or object appearance [21, 29, 55]), our framework integrates two complementary visual streams, facial and scene context, derived from the same video, enabling individual-source training while remaining applicable to real-world single-video inputs.

2.3. Flow Matching Models

Flow matching [17, 39] has recently gained attention as an efficient alternative to diffusion models, offering faster inference by following shorter and more direct generation trajectories. Recent works have applied flow matching to audio synthesis, separation, and enhancement [22, 29, 31, 35, 45, 58, 64], demonstrating its potential to produce high-quality, natural-sounding outputs. Previous non-generative, masking-based separation models often introduce artifacts (*e.g.*, spectral holes) as noted in [64], rendering the output unsuitable for downstream tasks like audio editing. We therefore adopt a generative flow-matching model for the CASS task, which effectively resolves this issue. To our knowledge, this is the first visually conditioned generative flow-matching approach to CASS, designed to yield perceptually natural and artifact-free separated audio suitable for cinematic production.

3. Methodology

We propose a framework for audio-visual cinematic audio source separation (AV-CASS). As shown in Fig. 2, it consists of a **Vision Extractor** that generates a fused representation $\boldsymbol{c}^V$ from input videos to condition the separation model. Next, a **flow-based generative model** performs source separation by mapping Gaussian noise to three distinct audio components: speech, sound effects, and music. Finally, to enable training without source-separated film data, we introduce a **training data synthesis pipeline** that

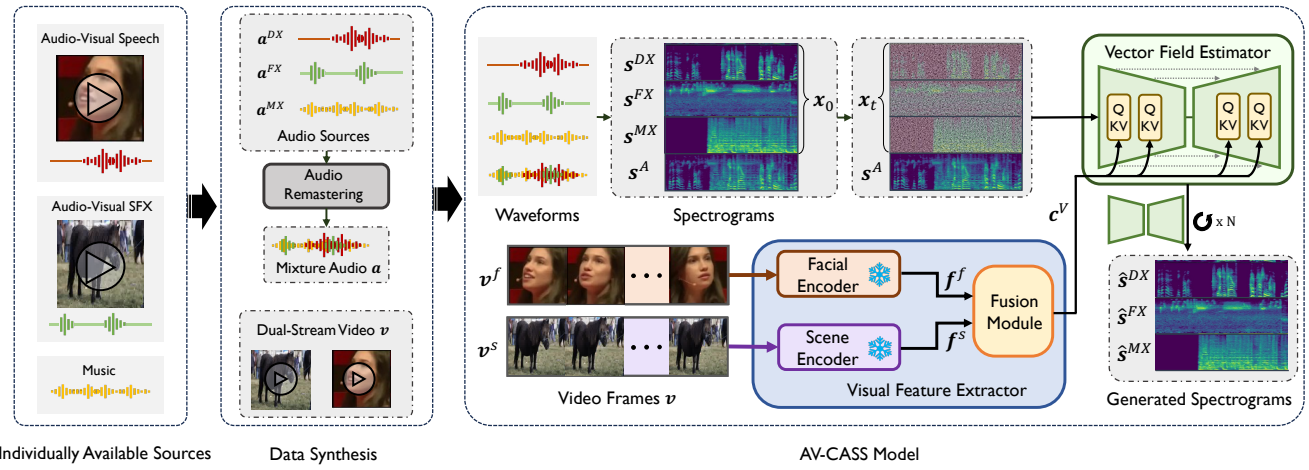

Figure 2. **Architecture of AV-CASS.** The fusion module integrates visual features from the facial and scene encoders into $\boldsymbol{c}^V$, which serves as a conditioning input along with a mixture audio $\boldsymbol{s}^A$ for the vector field estimator $\boldsymbol{u}_\theta$.

leverages individually available in-the-wild audio and video sources.

3.1. Problem Setup

Given a cinematic audio mixture $\boldsymbol{a}$ composed of three additive sources, speech $\boldsymbol{a}^{DX}$, sound effects $\boldsymbol{a}^{FX}$, and music $\boldsymbol{a}^{MX}$, the goal of CASS is to recover each source track from the mixture: $\boldsymbol{a} = \boldsymbol{a}^{DX} + \boldsymbol{a}^{FX} + \boldsymbol{a}^{MX}$. We formulate this task as a conditional generation problem, where the model is conditioned not only on the mixture audio but also on visual input $\boldsymbol{v}$ from the corresponding video. The video input contains two components: facial frames $\boldsymbol{v}^f$ and scene frames $\boldsymbol{v}^s$, which provide complementary cues for speech and sound effects, respectively.

3.2. Input Data for Training

To enable training without the need for source-separated cinematic data, we synthesize a training data by combining individual video and audio segments from diverse sources. For each training sample, we: (1) select speech clips from an audio-visual speech dataset to create the speech track $\boldsymbol{a}^{DX}$ and the corresponding video as face stream $\boldsymbol{v}^f$; (2) select sound effect clips from an audio-visual sound dataset to create the sound effects track $\boldsymbol{a}^{FX}$ and the corresponding video as scene stream $\boldsymbol{v}^s$; and (3) select background music clips from a music dataset to create the music track $\boldsymbol{a}^{MX}$. These source tracks are mixed to create the input mixture $\boldsymbol{a}$, while the video consists of two parallel streams, $\boldsymbol{v} = \{\boldsymbol{v}^f, \boldsymbol{v}^s\}$: a face stream and a scene stream that represent different semantic aspects of the sound. This strategy offers a diverse, controllable, and scalable source of training pairs with ground-truth supervision for all components. Details of the training data synthesis process are provided in Sec. 4. All audio waveforms are converted into their corresponding spectrograms where $\boldsymbol{s}^A$ denotes the mixture spectrogram, while $\boldsymbol{s}^{DX}$, $\boldsymbol{s}^{FX}$, and $\boldsymbol{s}^{MX}$ represent the individual source spectrograms. The video streams $\boldsymbol{v}^f$ and $\boldsymbol{v}^s$ are represented as sequences of frames.

3.3. AV-CASS Model

Our audio-visual cinematic source separation model (AV-CASS) consists of two main components: (1) visual feature extraction and fusion, and (2) conditional flow matching for source generation.

3.3.1. Visual Feature Encoding and Fusion

We extract visual features from two video streams: facial frames $\boldsymbol{v}^f$ and scene frames $\boldsymbol{v}^s$, using separate encoders suited to their semantic roles. The facial encoder, adopted from AVDiffuSS [37], is designed for lip-synced speech videos. The scene encoder, based on the CAVP model [41], captures temporally and semantically aligned features of sounding objects and events. Both encoders are frozen during training. The outputs are feature sequences $\boldsymbol{f}^f \in \mathbb{R}^{T_f \times D_f}$ and $\boldsymbol{f}^s \in \mathbb{R}^{T_s \times D_s}$, where T_f and T_s denote the number of frames, and D_f, D_s are feature dimensions. To fuse the representations, we project each stream into a shared feature space of dimension C using separate MLPs:

$$\boldsymbol{f}^f \to \mathbb{R}^{T_f \times C}, \quad \boldsymbol{f}^s \to \mathbb{R}^{T_s \times C}. \tag{1}$$

The resulting features are concatenated along the temporal axis and passed through a final fusion MLP to obtain the visual condition vector:

$$\boldsymbol{c}^V \in \mathbb{R}^{(T_f+T_s)\times C'}. \tag{2}$$

This visual condition is used to guide the audio generation process via cross-attention in the U-Net backbone.

3.3.2. Flow Matching for Multisource Separation

We adopt conditional flow matching [39] to model the conditional joint distribution of source spectrograms $p_1(\boldsymbol{x}) :=$

$p(\boldsymbol{s}^{DX}, \boldsymbol{s}^{FX}, \boldsymbol{s}^{MX})$ given the mixture spectrograms $\boldsymbol{s}^{A}$ and visual conditioning vector $\boldsymbol{c}^{V}$. Conditional flow matching defines a conditional mapping between the Gaussian noise distribution $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and the target joint distribution of source spectrograms $\boldsymbol{x}_1 \sim p_1(\boldsymbol{x})$. This mapping defines a time-varying probability density governed by an ordinary differential equation:

$$\mathrm{d}\boldsymbol{x}_t = \boldsymbol{u}_\theta(\boldsymbol{x}_t, t|\boldsymbol{c})\mathrm{d}t, \tag{3}$$

where $\boldsymbol{u}_\theta$ is a vector field estimator representing the gradient of the probability density w.r.t. time t at point $\boldsymbol{x}_t$; $\boldsymbol{x}_t$ is a point in the probability density space at time t; and $\boldsymbol{c}$ is conditioning variable includes mixture audio $\boldsymbol{s}^{A}$ and its visual context $\boldsymbol{c}^{V}$.

To construct this mapping, we train $\boldsymbol{u}_\theta$ to approximate a reference vector field $\boldsymbol{u}_t$, which constructs a probabilistic path between the noise distribution $p_0(\boldsymbol{x})$ and the target distribution $p_1(\boldsymbol{x})$, conditioned on $\boldsymbol{c}$. The vector field for a noise-data pair $(\boldsymbol{x}_0, \boldsymbol{x}_1)$ is defined as $\boldsymbol{u}_t = \boldsymbol{x}_1 - \boldsymbol{x}_0$. In [39], given a target distribution sample $\boldsymbol{x}_1$, the data point $\boldsymbol{x}_t$ for timestep t on the path is defined as:

$$\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \tag{4}$$

where $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and $t \in [0, 1]$. In practice, $\boldsymbol{u}_\theta$ is trained to approximate $\boldsymbol{u}_t$ by minimizing the following loss:

$$\mathcal{L} = \mathbb{E}_{t,\pi_1(\boldsymbol{x}_1),\pi_0(\boldsymbol{x}_0)} \|\boldsymbol{u}_\theta(\boldsymbol{x}_t, t|\boldsymbol{c}) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)\|_2^2. \tag{5}$$

Inspired by [17], we sample the timestep t from a logit-normal distribution, as this has been shown to enhance generation quality by placing more emphasis on the intermediate timesteps during training. In practice, we first sample a random variable z from a standard Gaussian distribution $z \sim \mathcal{N}(0, 1)$, then map it with a logistic function as follows:

$$t = \frac{1}{1 + e^{-z}}, z \sim \mathcal{N}(0, 1). \tag{6}$$

For the vector field estimator $\boldsymbol{u}_\theta$, we adopt a CNN-based U-Net architecture and three sources are concatenated along the channel dimension to form the input, as commonly done in diffusion-based image generation models [26, 51].

3.3.3. Inference

At inference, the trained conditional flow matching model generates separated source spectrograms from a mixture audio and its associated visual context. As outlined earlier, our training setup uses two video streams for a single audio mixture. Although this differs from the real-world one-video-one-audio setting, our approach remains effective for real-world cinematic content. Since different visual regions (*e.g.*, faces, environments, background elements) correspond to distinct sound sources, as shown in Fig. 3, we extract the facial regions as one stream for speech cues (processed by the Facial Encoder) and the full scene frames as another for sound effects (processed by the Scene Encoder). This design enables our model to process real-world samples without architectural changes. The model takes as input the mixture spectrogram $\boldsymbol{s}^{A}$ and fused visual condition $\boldsymbol{c}^{V}$ and outputs the separated components: speech, sound effects, and music.

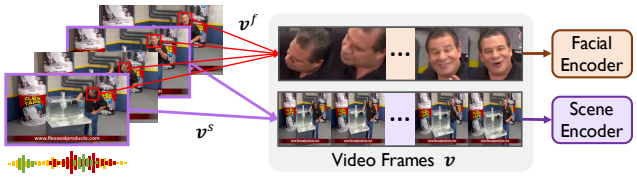

Figure 3. **Extraction of dual-stream visual inputs from a real-world cinematic video during inference.** Since no architectural changes are required, the AV-CASS model can be used with real-world cinematic videos for inference.

To formulate inference under the conditional flow matching framework, we follow the sampling procedure defined in [39]. Specifically, we initialize the sample $\boldsymbol{x}_0$ as Gaussian noise, $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and iteratively update it along the vector field predicted by vector field estimator $\boldsymbol{u}_\theta$, which is conditioned on both the mixture audio and visual context. We apply the forward Euler method to integrate the flow:

$$\boldsymbol{x}_{t+\eta} = \boldsymbol{x}_t + \eta\, \boldsymbol{u}_\theta(\boldsymbol{x}_t, t \mid \boldsymbol{s}^{A}, \boldsymbol{c}^{V}), \tag{7}$$

where $\eta = 1/N$ is the step size and N is the total number of sampling steps. The time variable t progresses from 0 to 1, and at each step, the vector field guides the sample closer to the data distribution of clean source spectrograms. After N steps, the final output $\boldsymbol{x}_1$ represents a concatenated output of three spectrograms, each corresponding to a separated source: $\{\hat{s}^{DX}, \hat{s}^{FX}, \hat{s}^{MX}\}$. These spectrograms are converted back to the waveform domain using inverse STFT.

4. Training Data Construction Pipeline

Training an audio-visual CASS model requires synchronized film video and isolated source audio tracks, which are rarely available for real-world cinematic content. To address this, we design a pipeline that synthetically pairs independently sourced audio and video from existing datasets to create realistic training pairs resembling cinematic data. This pipeline produces multimodal training data that mimics cinematic audio-visual patterns while preserving ground-truth alignment across all audio stems and video streams.

4.1. Individually Available Sources

We leverage two large-scale audio-visual datasets and one music dataset, each corresponding to one of the three target stems in CASS:

Speech (DX): We use LRS3 [2], a lip-synchronized video dataset with high-quality speech that may reflect cinematic dialogue through natural prosody and visual expressiveness. **Sound Effects (FX):** For ambient, object or action-driven sounds, we use VGGSound [6], an audio-visual dataset containing everyday events and objects. Unlike prior works [48, 60] using audio-only FSD50K, VGGSound provides aligned video for learning visually grounded effects. **Music (MX):** Since background music is typically not visually grounded, we follow standard practice and use the FMA dataset [13], which contains a wide variety of high-quality music.

4.2. Audio Preprocessing and Stream Synthesis

To ensure each sample contains a single, uncontaminated source track, we filter VGGSound and FMA with the SMAD model [30], removing all segments containing speech and music. After filtering, we obtain a total of 152K DX clips, $\sim$ 62K FX clips, and $\sim$ 49K MX samples.

Following the protocol in DnRv3 [60], we synthesize a cinematic audio by using the DX, FX, and MX tracks. For each track, we randomly sample short clips, concatenate them with overlapping transitions, and apply loudness normalization to meet cinematic mastering standards. The resulting tracks are then mixed by addition:

$$\boldsymbol{a}^{A} = \boldsymbol{a}^{DX} + \boldsymbol{a}^{FX} + \boldsymbol{a}^{MX}. \tag{8}$$

All audio is converted to mono and resampled at 16kHz. Statistics of the resulting training data are in *Appendix* C.

4.3. Visual Stream Synthesis

We extract the video clips corresponding to the DX and FX streams. Using the timestamps of each selected audio segment, we retrieve the facial video aligned with DX (from LRS3) and the scene video aligned with FX (from VGGSound). The MX stream has no associated visual input.

Each training sample contains a mixed audio stream with ground-truth sources and two video streams, *i.e.*, facial and scene, reflecting how speech and sound effects are visually grounded in real films. Although the dataset is synthetically constructed, it enables precise supervision for multimodal learning and aligns well with cinematic audio-visual conventions. As shown in the Sec. 5, our model trained on this data generalizes effectively to real-world movies, validating our training data pipeline.

5. Experiments

5.1. Experimental Setup

Baselines. We compare our method with existing CASS models, including MRX [48] and BandIt [59], as well as musical instrument separation models such as Hybrid Demucs [14], HT Demucs [52], and MSDM [44]. Beyond audio-only baselines, we also include the audio-visual sound separation model DAVIS-Flow [29] which is the current state-of-the-art model for audio-visual sound separation, to assess the contribution of visual conditioning and to highlight the fundamental differences between CASS and generic audio-visual separation tasks. All models are trained from scratch on the same dataset, using their original training configurations and appropriately modified to operate under the CASS setting for fair comparison.

Metrics. We use Fréchet Audio Distance (FAD) [32] and Kullback-Leibler divergence (KL) from AudioLDM [40] to measure distributional similarity between generated and real audio. We also report Perceptual Evaluation of Speech Quality (PESQ) [50] for speech and Scale-Invariant Signal-to-Distortion Ratio improvement (SI-SDRi) [36] in dB, following prior works [44, 48]. For FAD, KL, and SI-SDRi, we report averages across all three sources. In addition, we introduce a new metric, *Wrong Placement Ratio (WPR)*, to estimate the proportion of residual or misplaced components from other stems. WPR is computed using PANNs [34], a pretrained sound event detection model, and reflects stem-level separation quality without ground-truth references; lower values indicate better isolation. Details on metric calculations is in the *Appendix* E.

Training and Implementation Details. Before training the full audio-visual model, we apply audio-denoiser warm-up to stabilize optimization. Specifically, we first train the model using only the audio component of the synthetic audio-visual dataset described in Sec. 4. In practice, it improves training stability, accelerates convergence, and prevents the model from prematurely overfitting to visual cues in the early stages of training.

After warm-up, we train the full audio-visual model on the same synthetic dataset. Visual cues from both the facial and scene streams are introduced gradually using zero-initialized convolution layers, following ControlNet [65] strategy, while the video encoders remain frozen. This design preserves the stabilized audio representation while allowing the model to gradually integrate visual information.

We use the Adam optimizer [33] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a fixed learning rate of 10^{-4}, and no weight decay. Full audio-visual training runs for 600k steps with a batch size of 8 across four RTX 4090 GPUs. We use 128 sampling steps during evaluation. Additional implementation details and pseudo code are provided in the *Appendix* A.

5.2. Main Results

5.2.1. Evaluation on real-world samples

Subjective evaluation. Generalization to real-world movie samples is critical for CASS models. To evaluate our model, we randomly select clips from the Condensed Movies dataset [4], manually verifying that each contains all three

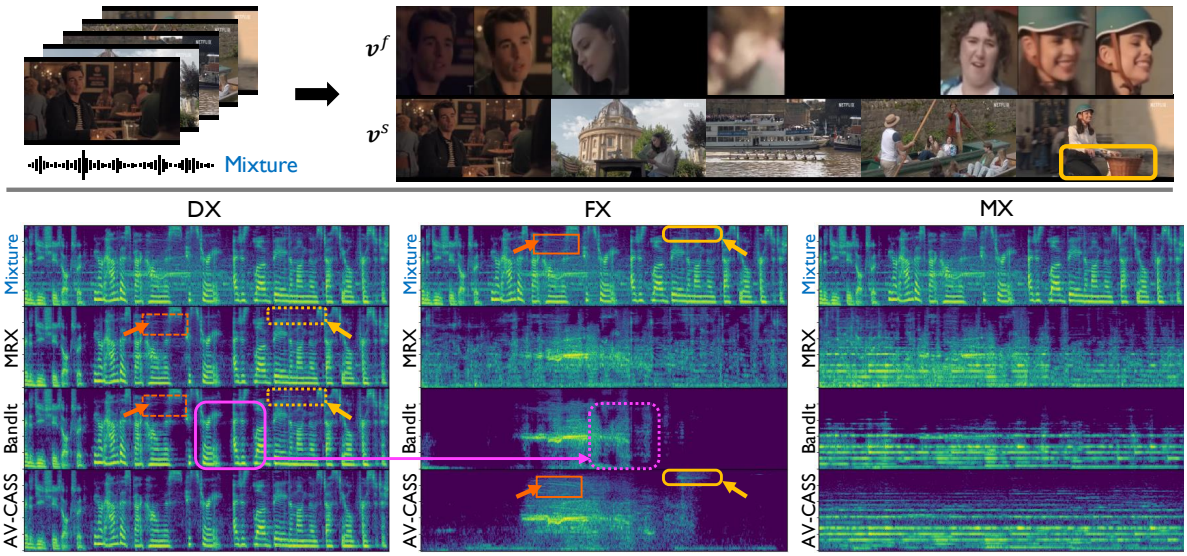

Figure 4. **Comparison of MRX, BandIt, and AV-CASS on a real-world movie sample.** Input video frames v^f and v^s are shown at the top, with the input audio spectrogram s^A placed for each stem. Yellow boxes highlight the bicycle bell, red boxes indicate cheering, and dotted boxes show elements misplaced in non-target stems. The dotted pink box in BandIt's FX shows speech artifacts. Best viewed when zoomed in. This sample can also be viewed in the supplementary video.

Method	MRX [48]	BandIt [59]	**AV-CASS (Ours)**
MOS (↑)	2.55 ± 0.10	3.78 ± 0.10	**4.13** $\pm$ **0.09**

Table 1. **MOS results of CASS models on real-world samples.** The scores are computed based on 95% confidence intervals.

Method	**DX** (↓)	**FX** (↓)	**MX** (↓)
MRX [48]	2.63	33.72	6.01
BandIt [59]	0.60	22.41	0.35
DAVIS-Flow [29]	5.88	**14.58**	35.94
AV-CASS (Ours)	**0.46**	19.81	**0.32**

Table 2. **Wrong Placement Ratio (WPR [%]) on real-world samples.** Lower is better.

target tracks. As this process requires extensive effort, we collected 30 samples. Since ground-truth separation data is unavailable, we conduct a subjective evaluation using mean opinion scores (MOS) through a human study, detailed in the *Appendix* D. Some of the movie samples used for MOS are available in the supplementary video. We strongly encourage readers to view real-world samples there, along with comparisons to existing CASS methods. The numerical results are provided in Tab. 1. Participants compared and evaluated the separated outputs for each target stem using two criteria: clarity of separation and completeness of target reconstruction. These criteria allows raters to evaluate both how well the model suppresses non-target sources and how fully it preserves the target source without loss of content or quality. Each separation was rated on a 5-point Likert scale, with 1 meaning "Poor" and 5 meaning "Excellent". As shown in Tab. 1, our model receives higher scores, reflecting user preference and demonstrating the naturalness and sound quality of the outputs. Overall, this result indicates generalization to real-world movie clips.

Objective evaluation. To assess AV-CASS on real-world movies, we report quantitative WPR across DX, FX, and MX stems. As shown in Tab. 2, our model achieves the lowest WPR for both DX and MX, and delivers competitive performance on FX, indicating strong separation fidelity for dialogue and music. While DAVIS-Flow [29] attains a lower WPR on FX, this is expected since it is specifically designed for generic object-centric sound separation; however, it performs poorly on the other tracks. Overall, the results demonstrate that AV-CASS achieves robust cross-track isolation in complex, in-the-wild cinematic audio, supporting the effectiveness of our formulation and synthetic training pipeline.

Qualitative result. We also provide a qualitative visual analysis (Fig. 4) illustrating residual cross-track components in existing methods, whereas our approach yields cleaner separated tracks. Yellow boxes in v^s and the spectrograms highlight the bicycle bell, and red boxes indicate cheering. Dotted boxes show elements incorrectly placed in non-target stems. Other methods retain these effects in the speech track, while ours correctly assigns them to FX. The dotted pink box in BandIt's FX shows speech artifacts, which AV-CASS avoids. This sample and more are shown in supplementary videos.

5.2.2. Evaluation on AVDnR

While real-world movie audio offers valuable qualitative insights, it does not provide clean ground-truth stems for quantitative evaluation. To enable a more controlled and comprehensive assessment, we construct a fully supervised

Method	A-V	FAD (↓)	KL (↓)	SI-SDRi (↑)	PESQ (↑)	WPR (↓)
Predictive Model						
Hybrid Demucs [14]	✗	2.05	1.03	13.57	2.16	5.24
HT Demucs [52]	✗	2.08	1.06	13.41	2.06	9.23
MRX [48]	✗	3.47	1.67	10.60	1.89	14.91
BandIt [59]	✗	2.15	1.14	**14.40**	2.15	4.65
Generative Model						
MSDM [44]	✗	2.90	2.90	11.63	2.12	5.65
DAVIS-Flow [29]	✓	5.94	1.64	9.25	1.96	12.14
AV-CASS (Ours)	✓	**0.84**	**0.93**	12.32	**2.26**	**1.84**
Ours (Audio-only)	✗	1.63	1.15	12.23	2.08	2.01

Table 3. **Results on AVDnR dataset (objective scores).** All models are trained on our training data. A-V indicates audio-visual.

Method	H-Demucs	HT Demucs	MRX	BandIt	MSDM	Ours
MOS (↑)	3.14 ± 0.15	3.01 ± 0.14	1.90 ± 0.13	3.12 ± 0.14	2.79 ± 0.14	**3.90 ± 0.13**

Table 4. **MOS results on AVDnR dataset.**

audio-visual test set, AVDnR, using the same data synthesis pipeline described in Sec. 4. We strictly partition all source clips into disjoint training and testing splits to avoid overlap. The final AVDnR benchmark contains 1K audio-visual samples, each 60 seconds long, providing a reliable testbed with complete ground-truth sources.

Objective evaluation. We train all methods on our training set and report results on AVDnR in Tab. 3. Methods are grouped into *predictive* and *generative* categories. As commonly observed [61, 64], predictive models which optimized with reconstruction or SNR-based losses tend to achieve higher SI-SDRi, reflecting their stronger alignment with SNR-based metrics. However, they often produce over-smoothed estimates that limit perceptual fidelity [38, 53]. In contrast, generative models focus on producing realistic samples and therefore excel on perceptual metrics. Within this context, AV-CASS achieves the best FAD, KL, PESQ, and WPR scores across all methods, indicating superior perceptual quality and cleaner cross-track separation.

Notably, AV-CASS outperforms the audio-visual baseline DAVIS-Flow [29], even though both methods use visual information. This suggests that performance in AV-CASS depends not only on using vision, but on the way of providing visual cues. By incorporating facial and scene streams in a dual-stream setup and adopting a multi-source formulation, AV-CASS receives source-specific visual context that is not available in the single-target design of DAVIS-Flow, leading to more consistent and reliable track disentanglement. Detailed per-track metrics are in the *Appendix* F.

Subjective evaluation. Since objective scores do not always capture perceived audio quality, especially for generative models that may introduce realistic details beyond the reference, we additionally conduct a user study on AVDnR. As shown in Tab. 4, AV-CASS achieves the highest MOS among all methods. This perceptual preference aligns with our strong FAD, KL, and PESQ performance (Tab. 3), indicating that AV-CASS produces outputs that listeners consistently find clearer and more natural. Together, these results confirm that the perceptual advantages of our model extend beyond objective measures and translate directly into improved user experience.

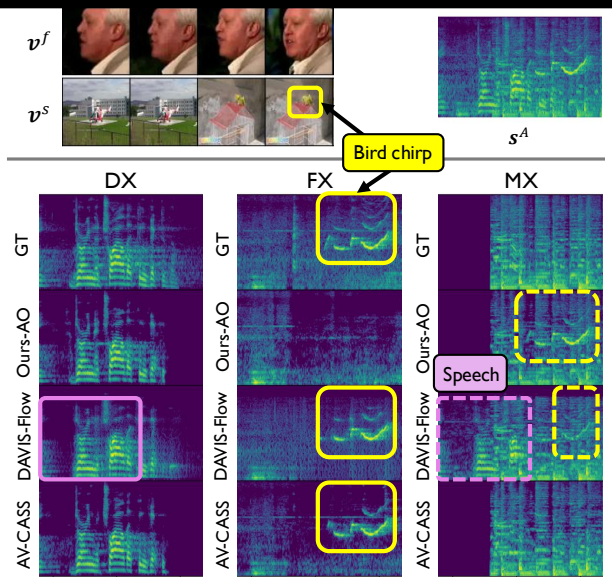

Figure 5. **Comparison of our audio-only model (Ours-AO), DAVIS-Flow [29], and our audio-visual model (AV-CASS) on a clip from the AVDnR testset.** The input video frames and the GT audio spectrograms are shown at the top. Yellow boxes highlight the bird chirping present in v^s and the FX tracks. Dotted boxes indicate misplaced segments. Better viewed when zoomed in.

Qualitative results. Fig. 5 compares separated spectrograms from our audio-only model (Ours-AO), DAVIS-Flow [29], and our audio-visual model (AV-CASS). As the yellow boxes indicate, both audio-visual models correctly separate bird chirping sound into FX, guided by the bird visible in v^s. In contrast, Ours-AO incorrectly places the chirping sound in the MX track. This shows that visual cues can guide correct source separation, resolving ambiguity about where the sound came from. While DAVIS-Flow separates FX correctly, it fails to isolate MX clearly because its design requires a corresponding visual input for every target source. This comparison highlights that AV-CASS effectively exploits visual information. Taken together with the results on real-world samples, the gap between AV-CASS and DAVIS-Flow further confirms that the CASS task differs from generic audio-visual sound separation and requires specialized architectures tailored to its multi-source separation demands, such as ours. Additional examples are shown in *Appendix* H.

5.2.3. Extendability towards audio-only separation

As discussed earlier, the CASS task is traditionally defined in the audio-only domain, with standard benchmarks such as DnRv2 [49] and DnRv3 [60]. Although our model is designed for audio-visual separation, it can also operate in

		FAD (↓)	KL (↓)	SI-SDRi (↑)	PESQ (↑)	WPR (↓)
DnRv2 [48]	Hybrid Demucs [14]	3.66	1.47	**9.19**	2.03	4.77
	HT Demucs [52]	3.72	1.37	8.59	1.95	9.14
	MRX [48]	5.01	1.77	7.48	1.78	16.56
	BandIt [59]	2.75	1.55	7.68	1.97	3.67
	MSDM [44]	6.25	1.54	9.09	**2.07**	5.33
	Ours (Audio-only)	**1.95**	**1.33**	8.10	1.93	**2.13**
DnRv3 [60]	Hybrid Demucs [14]	3.12	1.68	**10.62**	**1.89**	3.78
	HT Demucs [52]	3.17	**1.59**	9.92	1.83	8.49
	MRX [48]	5.02	2.26	8.94	1.65	16.64
	BandIt [59]	4.79	1.98	9.14	1.86	3.53
	MSDM [44]	5.51	1.75	9.19	1.65	4.49
	Ours (Audio-only)	**2.62**	1.66	9.36	1.86	**1.91**

Table 5. **Audio-only CASS results.** Metrics are averaged across three sources, except PESQ, which is evaluated only on the speech source. All models are trained on our dataset.

an audio-only setting, *e.g.*, when video frames are unavailable, by removing the visual encoder and cross-attention blocks in the U-Net. To evaluate this configuration, we train the audio-only variant on our dataset (as in previous experiments) and compare it with other methods on standard audio-only benchmarks and AVDnR. As shown in Tab. 3 and Tab. 5, our method is competitive with models specialized for audio-only CASS while showing clear superiority in FAD, indicating more natural and realistic outputs. Most importantly, our method achieves better WPR performance, indicating cleaner separation with less contamination. These results highlight the flexibility of our approach to support single-modality setups, though its primary scope remains solving CASS from an audio-visual perspective.

5.3. Ablation Study

Analysis on visual streams. We study the impact of visual stream by ablating facial and scene video streams. As shown in Tab. 6, adding either stream improves performance over the audio-only baseline, while using both yields the best results across all metrics, highlighting their complementary benefits and justifying our design choice for AV-CASS. We further analyze how each type of visual input affects the misplacement rates in Tab. 7. DX and FX columns clearly show that the corresponding visual inputs minimize misplaced segments: the lowest DX WPR is achieved with the facial stream, and the lowest FX WPR with the scene stream. By utilizing both visual cues, our model achieves the most balanced WPR performance across all stems and thus improves the overall perceptual performance. This fine-grained misplacement analysis provides additional evidence that visual cues not only improve signal quality metrics, but also substantially reduce semantic cross-contamination between tracks, a crucial property for practical cinematic audio applications. Taken together with the results in Tab. 6, these findings confirm that combining both visual streams offers the most consistent and reliable separation, yielding strong performance across metrics while minimizing residual cross-track contamination.

Method	v^f	v^s	FAD (↓)	KL (↓)	SI-SDRi (↑)	PESQ (↑)
Audio-only	✗	✗	1.63	1.15	12.23	2.08
+ Facial stream	✓	✗	0.91	1.00	12.13	2.21
+ Scene stream	✗	✓	0.87	1.00	12.27	2.24
+ Both (Ours)	✓	✓	**0.84**	**0.93**	**12.32**	**2.26**

Table 6. **Ablation results on visual streams.**

Method	v^f	v^s	DX	FX	MX
Audio-only	✗	✗	0.0265	4.6507	1.3443
+ Facial stream	✓	✗	**0.0164**	4.3310	**1.2028**
+ Scene stream	✗	✓	0.0289	**4.0939**	1.2282
+ Both (Ours)	✓	✓	0.0255	4.2854	1.2077

Table 7. **Ablation results on visual streams with Wrong Placement Ratio (WPR [%]) for each stem.** Lower is better.

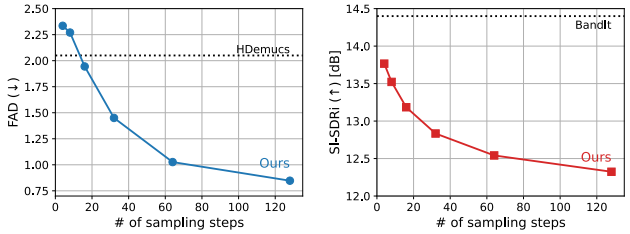

Figure 6. **Impact of the number of sampling steps.** Dotted lines denote Hybrid Demucs (second-best FAD) on the left and BandIt (highest SI-SDRi) on the right.

Impact of sampling step. Since we formulate CASS as a generative task, an important factor is the number of sampling steps N, which directly impacts performance. Fig. 6 shows FAD and SI-SDRi results on AVDnR as we vary $N \in \{4, 8, 16, 32, 64, 128\}$. FAD consistently improves with more steps, indicating better perceptual quality. Also, our model with only 32 steps already surpasses the second-best method, Hybrid Demucs [14], while also improving SI-SDRi. For our experiments, we use 128 steps to maximize perceptual quality, though N can be adjusted depending on the target metric.

6. Conclusion

In this work, we present the first audio-visual framework for cinematic audio source separation (CASS). By shifting from predictive to conditional generative modeling and leveraging multimodal cues, our method delivers high-quality, perceptually realistic separation. We develop a synthetic training pipeline that pairs in-the-wild audio and video, enabling training without cinematic datasets containing clean separated tracks. Trained solely on synthetic data, AV-CASS generalizes seamlessly to real cinematic content. Extensive experiments demonstrate its effectiveness on synthetic benchmark, real-world movie samples, and even standard audio-only CASS benchmarks. These results highlight the potential of an audio-visual perspective for building scalable, generalizable CASS models.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *Proc. Interspeech*, 2018. 1, 2

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition, 2018. 5

[3] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *ECCV*, 2020. 2

[4] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV*, 2020. 5, 2

[5] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *ICCV*, 2021. 2

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 5

[7] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. iquery: Instruments as queries for audio-visual sound separation. In *CVPR*, 2023. 1, 2

[8] Zhuo Chen, Yi Luo, and Nima Mesgarani. Deep attractor network for single-microphone speaker separation. In *ICASSP*, 2017. 1

[9] Xize Cheng, Siqi Zheng, Zehan Wang, Minghui Fang, Ziang Zhang, Rongjie Huang, Shengpeng Ji, Jialong Zuo, Tao Jin, and Zhou Zhao. Omnisep: Unified omni-modality sound separation with query-mixup. In *ICLR*, 2025. 2

[10] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *ACCV*, 2017. 2

[11] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Proc. Interspeech*, 2018. 1

[12] Soo-Whan Chung, Soyeon Choe, Joon Son Chung, and Hong-Goo Kang. Facefilter: Audio-visual speech separation using still images. In *Proc. Interspeech*, 2020. 2

[13] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017. 5

[14] Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021. 1, 5, 7, 8, 4

[15] Alexandre Défossez, Nicolas Usunier, Léon Bottou, and Francis Bach. Music source separation in the waveform domain, 2019. 1

[16] 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. In *Proc. ACM SIGGRAPH*, 2018. 1, 2

[17] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. ICML*, 2024. 1, 2, 4

[18] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *CVPR*, 2020. 2

[19] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *ICCV*, 2019. 2

[20] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, 2021. 1

[21] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *ECCV*, 2018. 1, 2

[22] Yiwei Guo, Chenpeng Du, Ziyang Ma, Xie Chen, and Kai Yu. Voiceflow: Efficient text-to-speech with rectified flow matching. In *ICASSP*, 2024. 1, 2

[23] Takuya Hasumi and Yusuke Fujita. Dnr-nonverbal: Cinematic audio source separation dataset containing non-verbal sounds. In *Proc. Interspeech*, 2025. 2

[24] John R. Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP*, 2016. 1

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 3

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 4

[27] Vincent Tao Hu, Wei Zhang, Meng Tang, Pascal Mettes, Deli Zhao, and Cees Snoek. Latent space editing in transformer-based flow matching. In *AAAI*, 2024. 1

[28] Chao Huang, Susan Liang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. DAVIS: High-quality audio-visual separation with generative diffusion models. In *ACCV*, 2024. 2

[29] Chao Huang, Susan Liang, Yapeng Tian, Anurag Kumar, and Chenliang Xu. High-quality sound separation across diverse categories via visually-guided generative modeling. *arXiv preprint arXiv:2509.22063*, 2025. 2, 5, 6, 7, 1, 4

[30] Yun-Ning Hung, Chih-Wei Wu, Iroro Orife, Aaron Hipple, William Wolcott, and Alexander Lerch. A large tv dataset for speech and music activity detection. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022. 5

[31] Chaeyoung Jung, Suyeon Lee, Ji-Hoon Kim, and Joon Son Chung. FlowAVSE: Efficient audio-visual speech enhancement with conditional flow matching. In *Proc. Interspeech*, 2024. 1, 2

[32] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. In *Proc. Interspeech*, 2019. 5, 3

[33] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[34] 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 Trans. on Audio, Speech, and Language Processing*, 2020. 5, 3, 4

[35] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. In *NeurIPS*, 2024. 1, 2

[36] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. SDR–half-baked or well done? In *ICASSP*, 2019. 5, 3

[37] Suyeon Lee, Chaeyoung Jung, Youngjoon Jang, Jaehun Kim, and Joon Son Chung. Seeing through the conversation: Audio-visual speech separation based on diffusion model. In *ICASSP*, 2024. 3, 1

[38] Jean-Marie Lemercier, Julius Richter, Simon Welker, and Timo Gerkmann. StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2023. 7

[39] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 1, 2, 3, 4

[40] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *Proc. ICML*, 2023. 5, 3

[41] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 2023. 3, 1

[42] Yi Luo and Jianwei Yu. Music source separation with band-split rnn. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2023. 2

[43] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 2022. 2

[44] Giorgio Mariani, Irene Tallini, Emilian Postolache, Michele Mancusi, Luca Cosmo, and Emanuele Rodolà. Multi-source diffusion models for simultaneous music generation and separation. In *ICLR*, 2024. 1, 5, 7, 8, 4

[45] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-TTS: A fast tts architecture with conditional flow matching. In *ICASSP*, 2024. 2

[46] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *CVPR*, 2018. 2

[47] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 2

[48] Darius Petermann, Gordon Wichern, Zhong-Qiu Wang, and Jonathan Le Roux. The cocktail fork problem: Three-stem audio separation for real-world soundtracks. In *ICASSP*, 2022. 1, 2, 5, 6, 7, 8, 4

[49] Darius Petermann, Gordon Wichern, Aswin Shanmugam Subramanian, Zhong-Qiu Wang, and Jonathan Le Roux. Tackling the cocktail fork problem for separation and transcription of real-world soundtracks. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2023. 2, 7

[50] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, 2001. 5, 4

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[52] Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP*, 2023. 1, 5, 7, 8, 4

[53] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 2022. 7

[54] Akira Takahashi, Shusuke Takahashi, and Yuki Mitsufuji. MMAudioSep: Taming video-to-audio generative model towards video/text-queried sound separation, 2025. 2

[55] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *ICLR*, 2021. 1, 2

[56] Efthymios Tzinis, Scott Wisdom, Tal Remez, and John R Hershey. Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation. In *ECCV*, 2022. 2

[57] Stefan Uhlich, Giorgio Fabbro, Masato Hirano, Shusuke Takahashi, Gordon Wichern, Jonathan Le Roux, Dipam Chakraborty, Sharada Mohanty, Kai Li, Yi Luo, Jianwei Yu, Rongzhi Gu, Roman Solovyev, Alexander Stempkovskiy, Tatiana Habruseva, Mikhail Sukhovei, and Yuki Mitsufuji. The sound demixing challenge 2023 – cinematic demixing track. *Trans. of the International Society for Music Information Retrieval*, 2024. 1

[58] 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. In *NeurIPS*, 2024. 1, 2

[59] Karn N. Watcharasupat, Chih-Wei Wu, Yiwei Ding, Iroro Orife, Aaron J. Hipple, Phillip A. Williams, Scott Kramer, Alexander Lerch, and William Wolcott. A generalized band-split neural network for cinematic audio source separation. *IEEE Open Journal of Signal Processing*, 2023. 1, 2, 5, 6, 7, 8, 4

[60] Karn N. Watcharasupat, Chih-Wei Wu, and Iroro Orife. Remastering divide and remaster: A cinematic audio source separation dataset with multilingual support. In *2024 IEEE 5th International Symposium on the Internet of Sounds*, 2024. 1, 2, 5, 7, 8

[61] Yutong Wen, Ke Chen, Prem Seetharaman, Oriol Nieto, Jiaqi Su, Rithesh Kumar, Minje Kim, Paris Smaragdis, Zeyu Jin, and Justin Salamon. PromptSep: Generative audio separation via multimodal prompting, 2025. 7

[62] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *ICASSP*, 2017. 1

[63] Yinfeng Yu and Shiyu Sun. Dgfnet: End-to-end audio-visual source separation based on dynamic gating fusion. In *Proc. ACM ICMR*, 2025. 2

[64] Yi Yuan, Xubo Liu, Haohe Liu, Mark D Plumbley, and Wenwu Wang. FlowSep: Language-queried sound separation with rectified flow matching. In *ICASSP*, 2025. 2, 7

[65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 5

[66] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, 2018. 2