Separating Invisible Sounds Toward
Universal Audiovisual Scene-Aware Sound Separation

Yiyang Su[1]*, Ali Vosoughi[2]*, Shijian Deng[3]*, Yapeng Tian[3], and Chenliang Xu[2]
[1]Michigan State University, suyiyan1@msu.edu
[2]University of Rochester, {mvosough@ur., chenliang.xu@}rochester.edu
[3]University of Texas at Dallas, {shijian.deng, yapeng.tian}@utdallas.edu

1. Introduction

Recent progress in audiovisual sound separation (AVSS) enables enhanced separation using aligned audiovisual cues [2, 12, 26, 27, 42–44, 53, 55]. Yet challenges arise when visual cues are absent, as in off-screen narrations or close-ups. We categorize sources as *visible sound* (in the visual scope) and *invisible sounds* (outside the visual scope).

Vital for sound separation is addressing invisible sounds. They are common in videos because most cameras have a limited field of view. Existing AVSS methods [42, 49, 53] can predict the difference between the sound mixture and the visible sounds as the invisible sound, yet they can not deal with multiple invisible sources.

We propose a novel audiovisual sound separation framework, **A**udio**V**isual **S**cene-**A**ware **Sep**aration (AVSA-Sep), which leverages video scene semantics as a substitute for visual cues. We contend that AVSS techniques leverage visual semantics for guiding sound separation. Hence, we incorporate audio semantics when visual semantics are absent.

As illustrated in Fig. 1, in our approach, we begin with audiovisual scene recognition to grasp scene-level semantics from the video. Next, an audiovisual separator predicts visible sounds, while a semantic-guided separator predicts invisible sounds. This scene-aware setup can handle both visible sounds and more than one invisible sound source.

Our contributions are threefold: 1) addressing invisible sounds, especially multiple invisible sounds, in audiovisual context; 2) introducing AVSA-Sep that leverages video scene semantics for visible and invisible sound separation; 3) incorporating semantic parsing into our framework, which helps to separate invisible sounds.

2. Related Work

Blind source separation (BSS). In audio signal processing, BSS methods untangle mixtures into source signals without additional cues like visuals. Classical approaches, such as Independent Component Analysis [18], Principal Component Analysis [17], and Non-negative Matrix Factorization [20, 35, 47], exploit statistical properties for source independence. Spatial audio-based methods [7, 8, 39] utilize location information. Deep-learning-based methods [16, 19, 48, 50] leverage deep networks to capture features. As shown in Tab. 1, while some BSS methods tackle multiple invisible sounds, audiovisual synergy is absent.

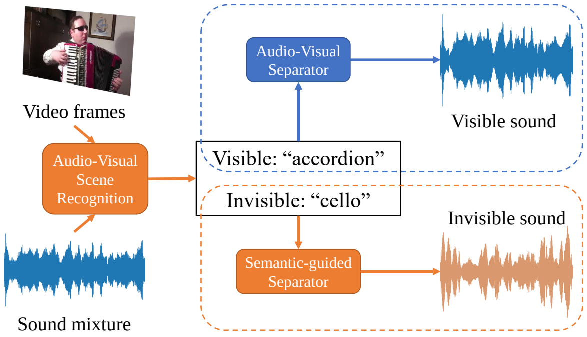

Figure 1. The AVSA-Sep framework. It predicts visible and invisible scenes from frames and sound mixture, then separates sounds. The audiovisual separator estimates visible sounds, while the semantic-guided one estimates invisible sounds. This separates both visible (e.g., blue waveform's accordion) and invisible (e.g., orange waveform's cello) sound sources.

Audiovisual sound separation. AVSS models have gained momentum since the inception of Sound of Pixels (SoP) [53]. Subsequent works [11, 43, 44, 49] have further harnessed audiovisual correspondences. Various aspects, such as motion [52], gestures [9, 38], natural language [37, 40], embodied AI [27, 28], object localization [12], visual grounding [41, 42], speech separation [13, 21, 33], and scene graphs [3, 4], have been incorporated. [34] considers offscreen sounds but is limited to one invisible sound. Similar to iQuery [5], we leverage semantic labels. While other AVSS methods handle up to one invisible sound, we address

*Equal contribution.

	BSS (e.g., [17])	**AVSS** (e.g., [53])	**AVSA-Sep** (Ours)
Uses visual cues	✗	✓	✓
2+ invisible sounds	Some	✗	✓

Table 1. BSS methods, such as RPCA [17], lack visual cues and sometimes handle up to 2 sounds. AV-Sep methods like SoP [53] address only one invisible sound, while our approach benefits from visual cues and can separate multiple invisible sounds.

the challenge of separating multiple invisible sounds.

Audiovisual scene understanding. Leveraging extensive audiovisual datasets [14, 22, 46] has driven a surge in audiovisual learning [1, 6, 10, 15, 23–25, 29–32, 45, 51, 54]. While these methods assume audiovisual correspondences, we incorporate semantics and representations of invisible sounds. Our approach further employs semantic information to guide the separation of invisible sounds.

3. Method

In the universal AVSS task, the goal is to recover individual sound sources $(S_1, S_2, \ldots, S_m)$ from visual frames $(I_1, I_2, \ldots, I_n)$ of n visible sounds and the sound mixture (S_{mix}) of m individual sounds. Traditional models assume a one-to-one correspondence ($m = n$), while our approach accommodates any m and n. For clarity, we assume each frame I_j corresponds to a sound S_i, as sounding object localization has been addressed by Tian *et al.* [41].

We introduce the AVSA-Sep framework to address this challenge. It leverages video scenes (frames or semantics) as intermediaries between input audio and separated sounds, as depicted in Fig. 1. The framework consists of two steps: 1) a semantic parser predicting scenes from visual frames and audio mixture and 2) sound separators separating sound components from the mixture conditioned on either visual frames or semantic labels. For clarity, we intentionally maintain simple architectures, anticipating that adapting newer AVSS baselines to our framework will yield improved performance.

3.1. Semantic-Guided Sound Separation

To address invisible sound sources, we suggest substituting semantic embeddings for visual features, shown in Fig. 2. This choice arises from their potential to guide sound separation, akin to visual features. To encode semantics, we introduce a label alignment network aligning semantic labels with visual features.

In the semantic alignment network, we encode a semantic label into a one-hot vector. A linear layer followed by a `sigmoid` activation aligns this vector with visual features, yielding the semantic label embedding f_s as the output.

The semantic alignment network introduces a semantic branch (*semantic-guided separator*) alongside the visual branch (*audiovisual separator*) in the sound separator. To ensure semantic embedding alignment with visual counterparts, we apply the same sound analysis network and sound synthesizer in both branches.

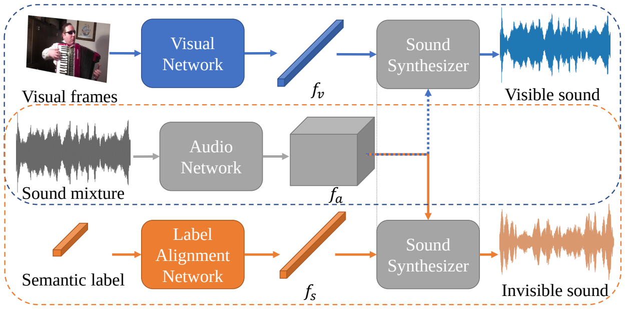

Figure 2. The sound separators of AVSA-Sep. The audiovisual separator is in blue the semantic-guided separator is in orange, with shared components in grey. The visual network generates visual features f_v, the label alignment network produces aligned semantic label feature embeddings f_s, and the audio network yields audio features f_a from the mixture spectrogram. Sound synthesizer predicts sound masks using audio and visual/semantic features and the same weights. Predicted masks combined with ISTFT recover predicted audios.

3.2. Audiovisual Scene Parser

To achieve audiovisual scene-aware separation, we introduce a semantic parser. It predicts semantic labels for sound sources through audiovisual classification. Then, we apply the audiovisual sound separator and the semantic-guided separator for distinct source types.

To cater to the semantic-guided separator, we predict both visible scene semantic labels L_{vis} and semantic labels for inaudible but audible scenes L_{inv}. Visible sounds are those that are both audible and visible, while invisible sounds are audible but not visible. Our semantic parser, shown in Fig. 3, comprises two main parts: the visible scene recognizer and the audible scene recognizer.

The visible parser includes a visual feature extractor, audio feature extractor, and fusion module merging audio and visual features. For video frames, a dilated ResNet-18 network generates k_r-channel visual features, pooled spatial-temporally. From audio mixture spectrograms, a VGGish network derives k_r-channel audio features through global pooling. The fusion module combines these features by summation. Fused features then transform semantic labels of visible scenes using a fully connected layer and `sigmoid` activation.

The audible scene recognizer derives a k_r-channel feature from the audio mixture's spectrogram. Through a fully connected layer and `sigmoid` activation, this feature transforms into semantic labels for audible scenes. It is important to note that the audio networks of the visible and audible scene recognizers have distinct weights due to differing

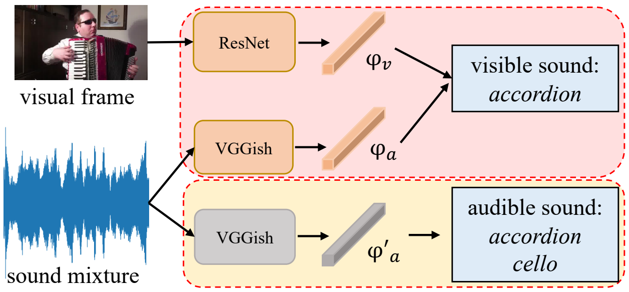

Figure 3. The figure shows the semantic parser. It employs a dilated ResNet-18 for visual features φ_v, and two VGGish networks for audio features φ_a and φ'_a. Predicted semantic labels for visible sounds come from the fused φ_v and φ_a, while audible sounds' labels originate from φ'_a.

tasks. The visible scene recognizer partners with the visual network to identify visible scenes, while the audible scene recognizer discerns both visible and invisible audio scenes.

3.3. Joint Sound Separation Training

During training, we independently train the semantic parser and sound separator, utilizing semantic labels for supervision and input, respectively. In inference, we merge the semantic parser and sound separator, enabling the whole model to predict both visible and invisible sounds from visual frames and sound mixture input.

The semantic parser is trained using a direct "mix-and-predict" approach. In each iteration, we mix sounds from randomly chosen videos, predict visible and audible sounds, and backpropagate loss. For AVSA-Sep and semantic parser training and evaluation, we adopt pipelines akin to those in Zhao *et al.* [53], with key modifications to optimize the joint training of the sound separator's two branches.

In each training and evaluation iteration, the sound separator processes a batch of k videos. For each video, we mix audios from $m - 1$ randomly selected video clips to form the audio mixture S_{mix}. The audiovisual separator utilizes frames from all m videos to predict individual sounds: $\tilde{S}_1^{\text{vis}}, \tilde{S}_2^{\text{vis}}, \dots, \tilde{S}_m^{\text{vis}}$. Similarly, the semantic-guided separator employs semantic labels of scenes from all m videos to predict individual sounds: $\tilde{S}_1^{\text{scn}}, \tilde{S}_2^{\text{scn}}, \dots, \tilde{S}_m^{\text{scn}}$.

The loss for visual and semantic masks relative to ground truth masks is given by

$$\sum_{i=1}^{k} \text{loss}(S_i, \tilde{S}_i^{\text{vis}}) \text{ and } \sum_{i=1}^{k} \text{loss}(S_i, \tilde{S}_i^{\text{scn}}), \tag{1}$$

respectively. The overall sound separation loss is defined as

$$\mathcal{L}_{\text{ss}} = \sum_{i=1}^{k} \lambda \cdot \text{loss}(S_i, \tilde{S}_i^{\text{vis}}) + (2 - \lambda) \cdot \text{loss}(S_i, \tilde{S}_i^{\text{scn}}), \tag{2}$$

where λ is a hyperparameter. Additionally, we include the triplet loss among ground truth masks as an anchor, visual/semantic masks as positive, and semantic/visual masks of other sources as negative items, with coefficient η as $\mathcal{L}_{\text{triplet}} = \sum_{i=1}^{k} \eta \cdot \text{triplet-loss}(S_i, \tilde{S}_i^{\text{vis}}, \tilde{S}_{-i}^{\text{scn}})) + \sum_{i=1}^{k} \eta \cdot \text{triplet-loss}(S_i, \tilde{S}_i^{\text{scn}}, \tilde{S}_{-i}^{\text{vis}}))$. The final loss becomes a combination of $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ss}} + \mathcal{L}_{\text{triplet}}$.

Dataset	Model	Visible		Invisible	
		SDR	SIR	SDR	SIR
MUSIC	Baseline [53]	-0.65	6.06	-1.92	0.24
	MP-Net [49]	**-0.19**	0.85	-1.73	-0.13
	AVSA-Sep (Ours)	-0.30	**6.90**	**-1.41**	**5.43**

Table 2. Comparison of our baseline (SoP [53]), MP-Net [49], and AVSA-Sep (ours) in terms of visible/invisible SDR/SIR on the MUSIC dataset under the challenging 3-sound setting.

These pipelines offer advantages such as joint weight updates during training and the combined evaluation of separation and scene recognition results against ground truth, yielding visual and semantic separation performance along with scene recognition performance.

4. Experiments

4.1. Experiment Setup

Implementation Details. We implement our framework based on SoP [53]. Following their work, we choose to use binary masks and log-scale spectrograms. Our experiments use 6-second audio clips with three evenly spread frames per video. Feature channels are set at $k_r = 512$, and separation loss scaling is $\lambda = 1.5$. Sound separation quality is quantified using SDR and SIR metrics from `mir_eval` [36] to assess audiovisual and semantic-guided separation outcomes.

Datasets. We trained and evaluated our model using the MUSIC dataset [53], comprising 500 user-uploaded videos highlighting 11 musical instruments. Each instrument category includes around 50 videos. For training and validation, we excluded duet videos (15%) lacking visible sounds. The dataset's clean and balanced nature is conducive to our task. We could generate artificial videos with invisible sounds for training and evaluation. Additionally, the dataset's accurate ground-truth scene labels are constructed from YouTube keyword queries, rendering it suitable for our purposes.

4.2. Quantitative Results

In our experiments, we generate mixtures involving 3 sounds to introduce challenging scenarios with the potential for multiple invisible sounds. When assessing visible sounds, we compare our audiovisual separator's performance with that of existing audiovisual sound separators.

As invisible sound has no visual cue, conventional audiovisual sound separators cannot utilize frames associated with invisible sounds. These separators need to leverage all other frames to generate an output. Therefore, we follow

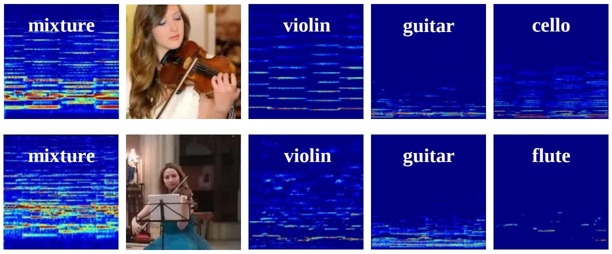

Figure 4. Two real-world trio videos. In both videos, only the violin is visible. Existing AVSS methods cannot separate the two invisible sounds. However, our approach can use semantic labels to separate them.

these steps: We execute the established sound separation methods as usual and acquire the predicted sound components $\tilde{S}_1^{\text{vis}}$, $\tilde{S}_2^{\text{vis}}$, and $\tilde{S}_3^{\text{vis}}$. Then, we subtract the predicted sound for other sources,

$$\tilde{S}_1^{\text{inv}} = S_{\text{mix}} - \tilde{S}_2^{\text{vis}} - \tilde{S}_3^{\text{vis}}, \tag{3}$$

and similarly for $\tilde{S}_2^{\text{inv}}$ and $\tilde{S}_3^{\text{inv}}$, where S_{mix} represents the input audio mixture.

This approach ensures that, during sound separation, audiovisual models disregard the associated frames for each sound, effectively treating both sounds as invisible. Consequently, the achieved metrics are comparable to those of the semantic-guided separator.

We present the results in Tab. 2. The results show that AVSA-Sep outperforms the baseline (SoP [53]) and MP-Net [49] in terms of both visible and invisible sound separation quality, as indicated by higher SDR or SIR values. This highlights the effectiveness of the proposed AVSA-Sep model in handling both visible and invisible sound sources.

4.3. Real-World Examples

We extend our evaluation beyond synthetic test cases to real-world video clips. To achieve this, we curate YouTube videos featuring 3 sounds and crop them to show only one of the musical instruments as visible. This approach allows us to gather real-world videos containing multiple invisible sounds. An illustrative instance of such a video, featuring one visible and two invisible sounds, is presented in Fig. 4.

In this example, our model uses the audiovisual separator to estimate visible sounds and the semantic-guided separator to estimate invisible sounds. Note that since AVSS methods assume all sounds are visible, they cannot separate any of the two invisible sounds.

Even though there is no ground-truth spectrogram for these videos, we can still see from predicted spectrograms that our model is capable of separating visible sounds and multiple invisible sounds.

Model	Visible		Invisible	
	SDR	SIR	SDR	SIR
AV-Sep / Baseline [53]	7.74	14.80	-	-
SG-Sep	-	-	6.23	12.80
AVSA-Sep on Train Set	10.73	17.67	9.72	16.22
AVSA-Sep w/ Train Frames	9.06	16.92	-	-
AVSA-Sep	7.91	14.81	9.01	16.45

Table 3. Ablation study results on the MUSIC dataset. SG-Sep refers to only the semantic-guided separator. [“On Train Set”: results evaluated on the training set; “w/ Train Frames”: test frames are replaced with training frames of the same category.]

4.4. Ablation Studies

Joint training. Given our dual-branch design, we explore the interactions between the two branches by training them separately. The results are summarized in Tab. 3.

Joint training enhances the performance of both the semantic-guided and audiovisual separators. This improvement is likely attributed to the use of complementary information from audiovisual modalities and semantic labels.

The visual branch *vs.* the semantic branch. The results in Tab. 3 indicate that the semantic branch outperforms the visual branch when both are present. One potential explanation is the semantic branch’s access to ground-truth labels, which the visual branch lacks. However, the semantic branch, without visual frames, cannot capture individual variations in videos. This is supported by our observation that while the semantic-guided separator achieves superior metrics on the test set, audiovisual separation performs better on the training set.

To investigate this further, we conduct an experiment where we replace the frames of test videos with frames from the training videos having the same scene label. The results, as summarized in Tab. 3, reveal a significant performance improvement when replacing the frames. This strongly supports the hypothesis that the underperformance of the visual branch is attributed to its inability to capture the semantic content in the frames of the test videos.

5. Conclusion and Limitations

Conclusion. We address the challenge of separating invisible sounds in AVSS and introduce a compatible framework with existing models. Our experiments confirm its capability to extend AVSS to multiple invisible sounds.

Limitations. While our approach handles multiple invisible sounds, exploring the treatment of multiple instances of the same semantic category could be valuable. Additionally, refining the semantic parser to enhance label prediction accuracy and potentially predict the counts of visible and invisible sounds represents another promising avenue.

References

[1] Triantafyllos Afouras, Yuki M Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. Self-supervised object detection from audio-visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10575–10586, 2022. 2

[2] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *European Conference on Computer Vision*, 2020. 1

[3] Moitreya Chatterjee, Narendra Ahuja, and Anoop Cherian. Learning audio-visual dynamics using scene graphs for audio source separation. *Advances in Neural Information Processing Systems*, 35:16975–16988, 2022. 1

[4] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1204–1213, 2021. 1

[5] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi. iquery: Instruments as queries for audio-visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14675–14686, 2023. 1

[6] Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision*, pages 431–448. Springer, 2022. 2

[7] Derry Fitzgerald, Antoine Liutkus, and Roland Badeau. Projection-based demixing of spatial audio. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1560–1572, 2016. 1

[8] Derry Fitzgerald, Antoine Liutkus, and Roland Badeau. Projet—spatial audio separation using projections. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 36–40. IEEE, 2016. 1

[9] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020. 1

[10] Junyu Gao, Mengyuan Chen, and Changsheng Xu. Collecting cross-modal presence-absence evidence for weakly-supervised audio-visual event perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18827–18836, 2023. 2

[11] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 1

[12] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3879–3888, 2019. 1

[13] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE, 2021. 1

[14] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 2

[15] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[16] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016. 1

[17] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 57–60. IEEE, 2012. 1, 2

[18] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000. 1

[19] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913, 2017. 1

[20] Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000. 1

[21] Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, and Kwanghoon Sohn. Looking into your speech: Learning cross-modal affinity for audio-visual speech separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1336–1345, 2021. 1

[22] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10284, 2021. 2

[23] Sangmin Lee, Hyung-Il Kim, and Yong Man Ro. Weakly paired associative learning for sound and image representations via bimodal associative memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10534–10543, 2022. 2

[24] Sangmin Lee, Sungjune Park, and Yong Man Ro. Audio-visual mismatch-aware video retrieval via association and adjustment. In *European Conference on Computer Vision*, pages 497–514. Springer, 2022. 2

[25] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-

visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309, 2023. 2

[26] Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Jinzheng Zhao, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Separate what you describe: Language-queried audio source separation. In *INTERSPEECH*, 2022. 1

[27] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 275–285, 2021. 1

[28] Sagnik Majumder and Kristen Grauman. Active audio-visual separation of dynamic sound sources. *arXiv preprint arXiv:2202.00850*, 2022. 1

[29] Otniel-Bogdan Mercea, Thomas Hummel, A Sophia Koepke, and Zeynep Akata. Temporal and cross-modal attention for audio-visual zero-shot learning. In *European Conference on Computer Vision*, pages 488–505. Springer, 2022. 2

[30] Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10553–10563, 2022. 2

[31] Himangi Mittal, Pedro Morgado, Unnat Jain, and Abhinav Gupta. Learning state-aware visual representations from audible interactions. *Advances in Neural Information Processing Systems*, 35:23765–23779, 2022. 2

[32] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. *Advances in Neural Information Processing Systems*, 35:34722–34733, 2022. 2

[33] Juan F. Montesinos, Venkatesh S. Kadandale, and Gloria Haro. Vovit: Low latency graph-based audio-visual voice sseparation transformer. In *European conference on computer vision*. Springer, 2022. 1

[34] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1

[35] Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE transactions on audio, speech, and language processing*, 18(3):550–563, 2009. 1

[36] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014. 3

[37] Akam Rahimi, Triantafyllos Afouras, and Andrew Zisserman. Reading to listen at the cocktail party: Multi-modal speech separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10493–10502, 2022. 1

[38] Tanzila Rahman, Mengyu Yang, and Leonid Sigal. Tribert: Human-centric audio-visual representation learning. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1

[39] Scott Rickard. The duet blind source separation algorithm. In *Blind speech separation*, pages 217–241. Springer, 2007. 1

[40] Reuben Tan, Arijit Ray, Andrea Burns, Bryan A Plummer, Justin Salamon, Oriol Nieto, Bryan Russell, and Kate Saenko. Language-guided audio-visual source separation via trimodal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10575–10584, 2023. 1

[41] Yapeng Tian, Di Hu, and Chenliang Xu. Co-learn sounding object visual grounding and visually indicated sound separation in a cycle. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 1, 2

[42] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2745–2754, 2021. 1

[43] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel P. W. Ellis, and John R. Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *International Conference on Learning Representations (ICLR) 2021*, 2021. 1

[44] Efthymios Tzinis, Scott Wisdom, Tal Remez, and John R Hershey. Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation. In *Proceedings of the European conference on computer vision (ECCV)*, 2022. 1

[45] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Sound and visual representation learning with multiple pretraining tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14616–14626, 2022. 2

[46] A Vedaldi, A Zisserman, H Chen, and W Xie. Vggsound: a large-scale audio-visual dataset. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2020. 2

[47] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE transactions on audio, speech, and language processing*, 15(3):1066–1074, 2007. 1

[48] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey. Alternative objective functions for deep clustering. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 686–690. IEEE, 2018. 1

[49] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 882–891, 2019. 1, 3, 4

[50] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017. 1

[51] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. 2

[52] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1735–1744, 2019. 1

[53] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 1, 2, 3, 4

[54] 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*, pages 386–403. Springer, 2022. 2

[55] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation using cascaded opponent filter network. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020. 1