iQuery: Instruments as Queries for Audio-Visual Sound Separation

Jiaben Chen[1], Renrui Zhang[2], Dongze Lian[3], Jiaqi Yang[4], Ziyao Zeng[4], Jianbo Shi[5]

[1]UC San Diego [2]The Chinese University of Hong Kong
[3]National University of Singapore [4]ShanghaiTech University [5]University of Pennsylvania

Abstract

Current audio-visual separation methods share a standard architecture design where an audio encoder-decoder network is fused with visual encoding features at the encoder bottleneck. This design confounds the learning of multi-modal feature encoding with robust sound decoding for audio separation. To generalize to a new instrument, one must fine-tune the entire visual and audio network for all musical instruments. We re-formulate the visual-sound separation task and propose Instruments as Queries (iQuery) with a flexible query expansion mechanism. Our approach ensures cross-modal consistency and cross-instrument disentanglement. We utilize "visually named" queries to initiate the learning of audio queries and use cross-modal attention to remove potential sound source interference at the estimated waveforms. To generalize to a new instrument or event class, drawing inspiration from the text-prompt design, we insert additional queries as audio prompts while freezing the attention mechanism. Experimental results on three benchmarks demonstrate that our iQuery improves audio-visual sound source separation performance. Code is available at https://github.com/JiabenChen/iQuery.

1. Introduction

Humans use multi-modal perception to understand complex activities. To mimic this skill, researchers have studied audio-visual learning [3, 17, 33] by exploiting the synchronization and correlation between auditory and visual information. In this paper, we focus on the sound source separation task, where we aim to identify and separate different sound components within a given sound mixture [60, 74]. Following the "Mix-and-Separate" framework [32, 34, 81], we learn to separate sounds by mixing multiple audio signals to generate an artificially complex auditory representation and then use it as a self-supervised task to separate individual sounds from the mixture. The works [26, 53, 89] showed that visually-guided sound separation is achievable by leveraging visual information of the sound source.

Prevalent architectures take a paradigm of a visual-conditioned encoder-decoder architecture [23, 26, 58, 88], where encoded features from audio and visual modalities are fused at the bottleneck for decoding to yield separated spectrogram masks. However, it is noticed that this design often creates a "muddy" sound and "cross-talk" that leaks from one instrument to another. To create a clean sound separation, one would like the audio-visual encoders to be (1) self-consistent within the music instrument and (2) contrasting across. One approach [27] added critic functions explicitly to enforce these properties. Another method [99] used a two-step process with the second motion-conditioned generation process to filter out unwanted cross-talks. We call these approaches decoder-centric.

Most recent works focus on addressing the "muddy" and "cross-talk" issue by improving fine details of audio-visual feature extraction: for example, adding human motion encoding as in [23, 88, 99], or cross-modality representations [58] via self-supervised learning. Once the feature representations are learned, the standard encoder-decoder FCN style segmentation is used as an afterthought. We consider these methods feature-centric. The standard designs have two limitations. First, it is hard to balance decoder-centric and feature-centric approaches that enforce a common goal of cross-modality consistency and cross-instrument contrast. Second, to learn a new musical instrument, one has to retrain the entire network via self-supervision.

To tackle these limitations, we propose a query-based sound separation framework, iQuery. We recast this problem from a query-based transformer segmentation view, where each query learns to segment one instrument, similar to visual segmentation [15, 16, 65, 78]. We treat each audio query as a learnable prototype that parametrically models one sound class. We fuse visual modality with audio by "visually naming" the audio query: using object detection to assign visual features to the corresponding audio query. Within the transformer decoder, the visually initialized queries interact with the audio features through cross-attention, thus ensuring cross-modality consistency. Self-

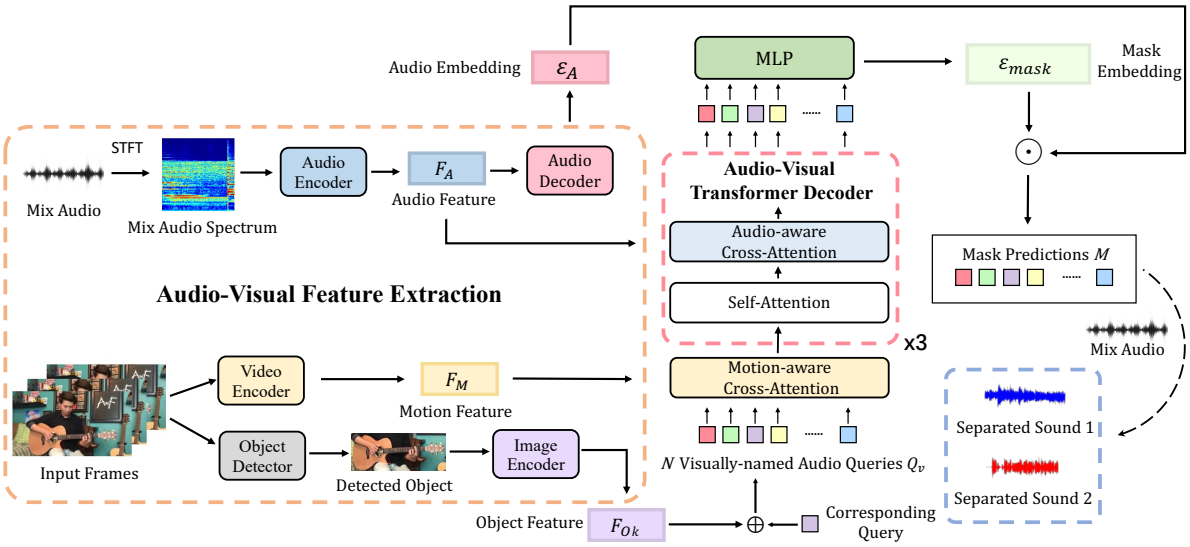


Figure 1. **Pipeline of iQuery.** Our system takes as input an audio mixture and its corresponding video frames, and disentangles separated sound sources for each video. Our pipeline consists of two main modules: an Audio-Visual Feature Extraction module which extracts audio, object, and motion features through three corresponding encoders, and an Audio-Visual Transformer module for sound separation. The query-based sound separation transformer has three key components: **1)** "visually-named" audio queries are initialized by extracted object features, **2)** cross-attention between the audio queries with static image features, dynamic motion features and audio features, **3)** self-attention between the learned audio queries to ensure cross-instrument contrast.

attention across the audio queries for different instruments implements a soft version of the cross-instrument contrast objective. With this design, we unify the feature-centric with the decoder-centric approach.

How do we achieve generalizability? Motivated by recent success in fine-tuning domain transfer with the text-prompt [28] and visual-prompt designs [7, 35, 41, 86], we adaptively insert the additional queries as audio prompts to accommodate new instruments. With the audio-prompt design, we freeze most of the transformer network parameters and only fine-tune the newly added query embedding layer. We conjecture that the learned prototype queries are instrument-dependent, while the cross/self-attention mechanism in the transformer is instrument-independent.

Our main contributions are:

- To the best of our knowledge, we are the first to study the audio-visual sound separation problem from a tunable query view to disentangle different sound sources explicitly through learnable audio prototypes in a mask transformer architecture.
- To generalize to a new sound class, we design an audio prompt for fine-tuning with most of the transformer architecture frozen.
- Extensive experiments and ablations verify the effectiveness of our core designs for disentanglement, demonstrating performance gain for audio-visual sound source separation on three benchmarks.

2. Related work

Audio-Visual Sound Source Separation. Recent years have witnessed promising results of audio-visual multi-modality joint learning [49, 62, 67, 75, 83] in domains like audio-visual sound source localization [4, 5, 14, 36, 55, 61, 63, 93], audio-visual event localization [68, 76, 77, 95] and sound synthesis from videos [25, 52, 54, 80, 97]. Sound source separation, a challenging classical problem, has been researched extensively in the audio signal processing area [11, 22, 37, 40]. A well-known example is the cocktail party problem [31, 48] in speech domain [1, 21]. Works have been proposed recently for tasks like speech separation [2, 27, 39, 51, 70], active sound separation [45, 46] and on-screen sound separation [25, 53, 71, 72]. Our work focuses on audio-visual sound separation. Recent audio-visual sound separation methods could be classified generally into two categories: feature-centric and decoder-centric as discussed in Sec. 1. Feature-centric methods exploit various ways for visual feature extraction selection to aid this multi-modality task. Some works consider frame-based appearance features (static frame features [24, 79, 89] or detected object regions [26, 66]) for extracting visual semantic cues (*e.g.*, instrument categories) to guide sound separation. [12, 13] adds embeddings from an audio-visual scene graph at the U-Net bottleneck to model the visual context of sound sources. Based on the assessment that motion signals

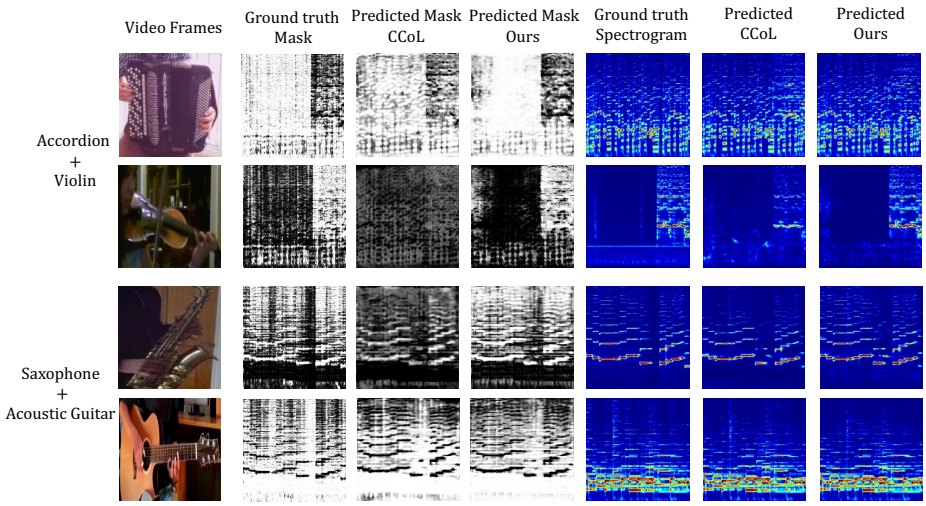


Figure 2. **Qualitative results on *MUSIC* test set.** The first column shows the mixed video frames, the second to the fourth columns compare our predicted spectrogram masks against masks yielded by state-of-the-art algorithm [66] and ground truth masks, and the fifth to the seventh columns visualize separated spectrograms. [66] produces blurry masks and contains unseparated components from another sound source, while our system successfully generates accurate mask and clean spectrograms as the ground truth.

could more tightly couple the moving sounding object with corresponding variations of sounds, recent approaches focus on including motion information into the pipeline (*e.g.*, optical flow [88], and human pose [23, 58]). Based on this, [94] proposes a framework to search for the optimal fusion strategy for multi-modal features. Decoder-centric methods explore prevention of "cross-talk" between the audio sources in the decoder stage. [99] designs a two-stage pipeline, where the second stage conducts a counterfactual synthesis through motion features to remove potentially leaked sound. The approach of [27] added critic functions explicitly to enforce cross-modal consistency and cross-instrument contrast.

Vision Transformers. Motivated by transformer's success in natural language processing [73], transformers were first introduced in computer vision for image classification as ViT [20]. Given the superior long-range modeling capacity, many follow-up works [47, 69, 82] have upgraded ViT to achieve higher performance and widely surpassed convolutional neural networks. Further, transformer-based models are adopted for various downstream tasks, such as 2D object detection [9, 91, 100], semantic/instance segmentation [65, 78, 92], 3D object detection [50, 85], shape recognition [84, 90] and video understanding [6, 42]. Particularly, following the pipeline from DETR [9], MaskFormer [16] and Mask2Former [15] represent each mask candidate as a learnable query and conduct parallel decoding for instance-level segmentation. However, only few approaches [39, 58, 71, 72, 99] have extended transformer for audio-visual sound separation fields. [58] adopts a BERT [18] architecture to learn visual, pose, and audio feature representations. [99] designs an audio-motion transformer to refine sound separation results through audio-motion feature fusion. These methods focus mainly on learning better contextualized multi-modality representations through an encoder transformer. In contrast, our mask transformer-based network focuses on the entire process of visual-audio separation task. We disentangle different sound sources through independent learnable query prototypes and segment each time-frequency region on the spectrogram via mask prediction in an end-to-end fashion.

3. Method

We first describe the formulation of the audio-visual sound separation task and introduce our pipeline iQuery briefly in Sec. 3.1. Then we introduce networks for learning representations from visual and audio modalities in Sec. 3.2 and our proposed cross-modality cross-attention transformer architecture for visual sound separation in Sec. 3.3. Finally, we introduce our adaptive query fine-tuning strategy through designs of flexible tunable queries in Sec. 3.4.

3.1. Overview

As mentioned before, our goal is to disentangle the audio mixture concerning its corresponding sound sources in the given mixture by using so-called queries. Following previous works [21, 89], we adopt a commonly used "Mix-and-Separate" self-supervised source separation procedure. Given K video clips with accompanying audio signal: $\{(V_k, s_k(t))\}_{k\in[1,K]}$, we create a sound mixture:

$s_{mix}(t) = \sum_{k=1}^{K} s_k(t)$ as training data. Our disentanglement goal is to separate sounds $s_k(t)$ from $s_{mix}(t)$ for sound sources in V_k, respectively. The pipeline, as illustrated in Fig. 1, is mainly composed of two components: an Audio-Visual Feature Extraction module and a Mask Transformer-based Sound Separation module. First, in the feature extraction module, the object detector & image encoder, and video encoder extract object-level visual features and motion features from video clip V_k. The audio network yields an audio feature and an audio embedding from the given sound mixture $s_{mix}(t)$. After that, a cross-modal transformer decoder attends to visual and audio features and outputs audio mask embeddings, which are further combined with audio embeddings for sound separation.

3.2. Audio-Visual Feature Extraction

Object Detector & Image Encoder. To initialize learning of audio queries, we assign object-level visual appearance features to the corresponding queries, to create "visually named" queries. In the implementation, following [26], we use a Faster R-CNN object detector with ResNet-101 backbone. For frames in a given video clip V_k, the object detector is utilized to acquire the detected objects set O_k. After that, we adopt a pre-trained ResNet-18 similar to [66], followed by a linear layer and max pooling to yield object-level features $F_{O_k} \in \mathbb{R}^{C_O}$, where C_O denotes channel dimension of object features.

Video Encoder. The video encoder maps the video frames from $V_k \in \mathbb{R}^{3\times T_k\times H_k\times W_k}$ into a motion feature representation. In contrast with previous motion representations [23, 58, 88, 99], we use self-supervised video representation obtained from a 3D video encoder of I3D [10] pre-trained by FAME [19]. The model is pre-trained contrastively to concentrate on moving foregrounds. Finally, a spatial pooling is applied to obtain motion embedding $F_{M_k} \in \mathbb{R}^{C_M\times T'_k}$, where C_M denotes the dimension of the motion feature.

Audio Network. The audio network takes the form of skip-connected U-Net style architectures [59] following [26, 66, 89]. Given the input audio mixture $s_{mix}(t)$, we first apply a Short-Time Fourier Transform (STFT) [30] to convert the raw waveform to a 2D Time-Frequency spectrogram representation $S_{mix} \in \mathbb{R}^{F\times T}$, which is then fed into the U-Net encoder to obtain an audio feature map $F_A \in \mathbb{R}^{C_A\times \frac{F}{S}\times \frac{T}{S}}$ (C_A denotes the number of channels and S denotes stride of audio feature map) at the bottleneck. A U-Net decoder gradually upsamples the audio features to yield audio embeddings $\varepsilon_A \in \mathbb{R}^{C_\varepsilon\times F\times T}$ (C_ε denotes the dimension of audio embeddings), which is combined further with the transformer mask embeddings to generate the separated sound spectrogram mask M_k.

3.3. Audio-Visual Transformer

Our cross-modality sound separation transformer contains the transformer decoder [73] with N queries (*i.e.*, learnable prototypes), and utilizes the extracted object features F_{O_k}, motion embeddings F_{M_k} and audio features F_A to yield N mask embeddings $\varepsilon_{mask} \in \mathbb{R}^{C_\varepsilon\times N}$ for spectrogram mask prediction of separated sound $s_k(t)$, where N denotes maximum of the pre-defined instrument types.

Audio query prototypes. We denote audio queries as $Q \in \mathbb{R}^{C_Q\times N}$ to represent different instruments, which are initialized by "visually naming" audio queries. Specifically, "visually naming" means that we assign object features F_{O_k} to the corresponding query in Q with element-wise addition to yield "visually-named" queries Q_v, which are then fed into the transformer decoder cross-attention layers.

Cross-attention layers. In the decoder, we stack one motion-aware decoder layer and three audio-aware decoder layers. The "visually-named" queries Q_v first interact temporally with motion features F_{M_k} in the motion-aware decoder layer with motion cross-attention by $\mathrm{Attention}(Q_v, F_{M_k}, F_{M_k})$. This is followed by an FFN to generate the motion-decoded queries Q', which are then fed into three audio-aware decoder layers to adaptively interact with audio features F_A, each of which consists of a self-attention, an audio cross-attention computed by $\mathrm{Attention}(Q', F_A, F_A)$, and an FFN. The output N audio segmentation embeddings $\varepsilon_Q \in \mathbb{R}^{C_Q\times N}$ is computed by

$$\varepsilon_Q = \mathrm{AudioDecoder}_{\times 3}(Q', F_A, F_A), \tag{1}$$

where $\mathrm{AudioDecoder}$ stands for our audio-aware decoder layer. Similar to [9, 16], the decoder generates all audio segmentation embeddings parallelly.

Separated mask prediction. Through the above decoder, the N audio segmentation embeddings ε_Q are converted to N mask embeddings $\varepsilon_{mask} \in \mathbb{R}^{C_\varepsilon\times N}$ through a MLP with two hidden layers, where dimension C_ε is identical to dimension of audio embeddings $\varepsilon_A \in \mathbb{R}^{C_\varepsilon\times F\times T}$. Then each predicted mask $M_k \in \mathbb{R}^{F\times T}$ of the separated sound spectrogram is generated by a dot-product between the corresponding mask embedding in ε_{mask} and audio embedding ε_A from the audio decoder. Finally, we multiply the sound mixture spectrogram S_{mix} and the predicted mask M_k to disentangle sound spectrogram S_k for sound $s_k(t)$ by

$$S_k = S_{mix} \odot M_k, \tag{2}$$

where $\odot$ denotes the element-wise multiplication operator. Ultimately, separated sound signal $s_k(t)$ is produced by applying inverse STFT to the separated spectrogram S_k.

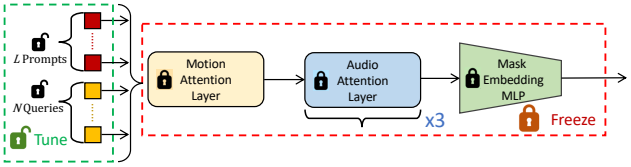


Figure 3. **Audio prompts design.** To generalize to new types of instruments/event classes, we propose to insert additional queries (*audio prompts*) to learn new audio prototypes for unseen classes. With this design, we only fine-tune the query embedding layer while keeping all the other parts of transformer backbone frozen.

Training objective. Following [26, 89], we set our training objective as optimizing spectrogram masks. The ground truth ratio mask M_k^{GT} of k-th video is calculated as follows,

$$M_k^{GT}(t,f) = \frac{S_k(t,f)}{S_{mix}(t,f)}, \tag{3}$$

where (t, f) denotes time-frequency coordinates. We adopt per-pixel $L1$ loss [87] to optimize the overall sound separation network, sound separation loss L_{sep} is defined as,

$$L_{sep} = \sum_{k=1}^{K} ||M_k - M_k^{GT}||_1, \tag{4}$$

where K denotes number of mixed sounds in S_{mix}.

3.4. Tunable Queries as Audio Prompts

With the flexible design of tunable queries as learnable prototypes, our pipeline is more friendly to generalizing to new types of instruments. Unlike previous methods that need to finetune the entire mask generation U-Net, we could insert additional queries(*i.e.*, audio prompts) for the new instruments. Such a method enables us only need to fine-tune the query embedding layer for learning new audio query prototypes in Sec. 3.3 of our transformer architecture while keeping all cross-attention layers frozen (see Fig.3). Specifically, we add L new audio prompts $P \in \mathbb{R}^{C_Q \times L}$ to original pre-trained audio queries $Q \in \mathbb{R}^{C_Q \times N}$, then the query embedding layer for the prompted learnable prototypes $Q_{prompted} \in \mathbb{R}^{C_Q \times (N+L)}$ is the only layer learnable in our transformer decoder, while keeping the transformer backbone frozen.

4. Experiments

4.1. Experimental Settings

Datasets. We perform experiments on three widely-used datasets: *MUSIC* [89], *MUSIC-21* [88], and *Audio-Visual Event (AVE)* [29, 68]. *MUSIC* dataset spans 11 musical instrument categories: accordion, acoustic guitar, cello, clarinet, erhu, flute, saxophone, trumpet, tuba, violin, and xylophone. This dataset is relatively clean, and sound sources are always within the scene, collected for the audio-visual sound separation task. We utilize 503 online available solo videos and split them into training/validation/testing sets with 453/25/25 videos from 11 different categories, respectively, following same settings as [66]. *MUSIC-21* dataset [88] is an enlarged version of *MUSIC* [89], which contains 10 more common instrument categories: bagpipe, banjo, bassoon, congas, drum, electric bass, guzheng, piano, pipa, and ukulele. We utilize 1,092 available solo videos and split them into train/test sets with 894/198 videos respectively from 21 different categories. Note that we follow the same training/testing split as [23, 99]. *AVE* dataset is a general audio-visual learning dataset, covering 28 event classes such as animal behaviors, vehicles, and human activities. We follow the same setting as [99], and utilize 4143 videos from AVE [68] dataset.

Baselines. For *MUSIC* dataset, we compare our method with four recent methods for sound separation. *NMF-MFCC* [64] is a non-learnable audio-only method, we consider reporting this result from [26, 58] on *MUSIC* test set. We also compare with two representative audio-visual sound separation baselines: *Sound-of-Pixels* [89] and *Co-Separation* [26]. We retrained these two methods with the same training data and split them as ours for a fair comparison. Finally, we compare our approach with a most recent publicly-available baseline *CCoL* [66], which has the same training setting as ours. For *MUSIC-21* dataset, we compare our method with six recently proposed approaches: *Sound-of-Pixels* [89], *Co-Separation* [26], *Sound-of-Motions* [88], *Music Gesture* [23], *TriBERT* [58] and *AMnet* [99]. For [58], since 12.27% of the training samples are missing in their given training split, we consider their reported result as a baseline comparison. Finally, for *AVE* dataset, we compare our method with six state-of-the-art methods. Since we conduct our experiments with the same setting as *AMnet* [99], we report results from [99] for *Multisensory* [53], *Sound-of-Pixels* [89], *Sound-of-Motions* [88], *Minus-Plus* [79], *Cascaded Opponent Filter* [98] as baseline comparisons.

Evaluation metrics. The sound separation performance is evaluated by the popular adopted mir_eval library [57] in terms of standard metrics: Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), and Signal to Artifact Ratio (SAR). SDR measures the combination of interference and artifacts, SIR measures interference, and SAR measures artifacts. For all three metrics, a higher value indicates better results.

Implementation Details. For *MUSIC* [89] and *MUSIC-21* [88] datasets, we sub-sample the audio at 11kHz, and each

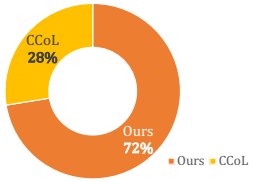


Figure 4. **Human evaluation results** for sound source separation on mixtures of different instrument types. Our system is able to separate sounds with better actual perceptual quality.

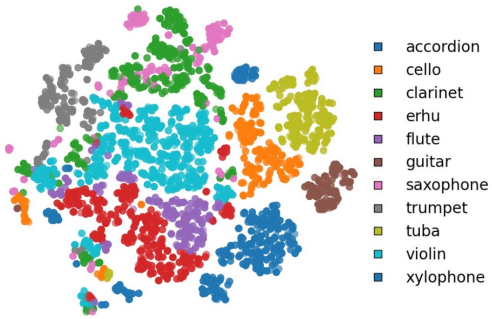


Figure 5. **Visualization of audio query embeddings** with t-SNE, different instrument categories are color-coded. Our audio queries have learned to cluster by different classes of sound.

audio sample is approximately 6 seconds. STFT is applied using a Hann window size of 1022 and a hop length of 256, yielding a 512×256 Time-Frequency audio representation. It is then re-sampled on a log-frequency scale to obtain a magnitude spectrogram with $T, F = 256$. Detected objects in frames are resized to 256×256 and randomly cropped to the size of 224×224. We set the video frame rate as 1 FPS, and randomly-selected three frames as input for the object detector. While for *AVE* [68] dataset, audio signal is sub-sampled at 22kHz, and we use the full frame rate(29.97 FPS). Other settings are the same as *MUSIC* except STFT hop length is set as 184, following [99].

For *MUSIC* dataset [89], we use the Faster R-CNN object detector pre-trained by [26] on Open Images [38]. For *MUSIC-21* [88] and *AVE* [68] datasets, since additional musical and general classes are not covered for this object detector, we adopt a pre-trained Detic detector [96] based on CLIP [56] to detect the 10 more instruments in *MUSIC-21* dataset [88] and 28 event classes in *AVE* dataset [68].

We utilize 8 heads for all attention modules and select the maximum N objects (number of queries) as 15, 25, and 30 for MUSIC, MUSIC-21 and AVE. The video encoder [19] and the object detector is pre-trained and kept frozen during training and inference. The multi-layer perception (MLP) for separated mask prediction has 2 hidden layers of 256 channels following [16]. Audio feature F_A, motion feature

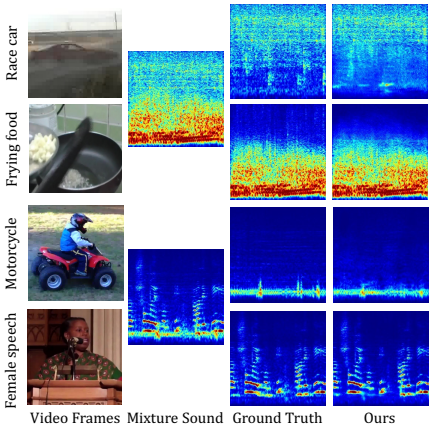


Figure 6. **Qualitative results on AVE test dataset.** Beyond restricted musical instruments, our model is also able to handle general sound separation tasks (**e.g.** sounds of galloping race car and frying food on the first two rows; sounds of driving motorcycles and speeches on the last two rows).

F_M, object feature F_O, and audio queries Q have a channel dimension of 256. And we set the channel dimension of both audio embeddings ε_A and mask embeddings ε_M as 32. The epoch number is 80, and batch size is set to 8. We use AdamW [43] for the mask transformer with a weight decay of 10^{-4} and Adam for all other networks as optimizer selection. The learning rate of the transformer is set as 10^{-4} and decreases by multiplying 0.1 at 60-th epoch. We set the learning rate for other networks as 10^{-4}, decreased by multiplying 0.1 at 30-th and 50-th epoch, respectively. Training is conducted on 8 NVIDIA TITAN V GPUs.

4.2. Audio-Visual Sound Source Separation

Quantitative evaluation. Table. 1 demonstrates quantitative results for sound separation results against state-of-the-art methods on *MUSIC* dataset [89]. Our method outperforms baseline models in separation accuracy measured by all evaluation metrics. Our method outperforms the most recent publicly available state-of-the-art algorithm [66] by 3.43 dB in terms of SDR score. Regarding quantitative results on *MUSIC21* dataset [88], we demonstrate the performance comparison in Table. 2. Again, our method outperforms baseline models in terms of SDR metric. Performance on the previous two datasets demonstrate our model's ability to disentangle musical sounds. To further verify the scalability of our proposed method to general audio-source separation problems, we perform quantitative comparisons on *AVE* dataset in Table. 3. As is demonstrated, we surpass the state-of-the-art algorithm [99] by 1.31 dB in terms of SDR score. *AVE* is

Methods	SDR↑	SIR↑	SAR↑
NMF-MFCC [64]	0.92	5.68	6.84
Sound-of-Pixels [89]	4.23	9.39	9.85
Co-Separation [26]	6.54	11.37	9.46
CCoL [66]	7.74	13.22	11.54
iQuery (Ours)	**11.17**	**15.84**	**14.27**

Table 1. **Audio-visual sound separation results on *MUSIC*.** Best results in **bold** and second-best results in Blue.

Methods	SDR↑	SIR↑	SAR↑
Sound-of-Pixels [89]*	7.52	13.01	11.53
Co-Separation [26]*	7.64	13.80	11.30
Sound-of-Motions [88]*	8.31	14.82	13.11
Music Gesture [23]*	10.12	15.81	-
TriBERT [58]	10.09	17.45	12.80
AMnet [99]*	11.08	**18.00**	13.22
iQuery (Ours)	**11.12**	15.98	**14.16**

Table 2. **Audio-visual sound separation results on *MUSIC-21*.** The results noted by * are obtained from [23, 99].

a general dataset containing scenes like male and female speeches, animal sounds, and vehicle sounds. This clearly shows our model's adaptivity to more general problems of sound source separation.

Qualitative evaluation. Fig. 2 illustrates qualitative sound separation results on *MUSIC* dataset. It can be seen that our method disentangles sound sources cleaner and more accurately, with less "muddy" sound. Fig. 6 provides additional qualitative examples on *AVE* dataset, and this again illustrates our model's good performance on general sound source separation cases. Both qualitative and quantitative results verify the superiority of our designed sound query-based segmentation pipeline iQuery.

Human evaluation. Our quantitative evaluation shows the superiority of our model compared with baseline models, however, studies [8] have shown that audio separation quality could not be truthfully determined purely by the widely used mir_eval [57] metrics. Due to this reason, we further conduct a subjective human evaluation to study the actual perceptual quality of sound-separation results. Specifically, we compare the sound separation result of our model and the publicly available best baseline model [66] on *MUSIC* [89]. We collected 50 testing samples for all 11 classes from the test set, and each testing sample contains separated sounds with a length of 6 seconds predicted by our model and baseline [66] for the same sound mixture. Ground truth sound is also provided for each sample as a reference. The experiment is conducted by 40 participants separately. For each participant, the orders of our model and baseline [66] are randomly shuffled, and we ask the participant to answer "*Which sound separation result is more close to the ground truth audio?*" for each sample. Statistical results are shown in Fig. 4. Notably, our method significantly surpasses the compared baseline with a winning rate of 72.45%. This additionally demonstrate the better actual perceptual performance of our model.

Methods	SDR↑	SIR↑	SAR↑
Multisensory [53]*	0.84	3.44	6.69
Sound-of-Pixels [89]*	1.21	7.08	6.84
Sound-of-Motions [88]*	1.48	7.41	7.39
Minus-Plus [79]*	1.96	7.95	8.08
Cascaded Filter [98]*	2.68	8.18	8.48
AMnet [99]*	3.71	**9.15**	11.00
iQuery (Ours)	**5.02**	8.21	**12.32**

Table 3. **Audio-visual sound separation results on *AVE*.** The results noted by * are obtained from [99].

Methods	SDR↑	SIR↑	SAR↑
Sound-of-Pixels [89]	4.11	8.17	9.84
Co-Separation [26]	5.37	9.85	8.72
CCoL [66]	6.74	**11.94**	10.22
iQuery (Ours)	**8.04**	11.60	**13.21**

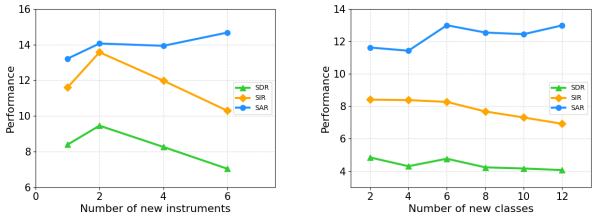
Table 4. **Fine-tuning sound separation performance comparison.** All methods are pretrained on *MUSIC* dataset without one particular instrument and then fine-tuned on this new data. Baseline models are tuned with whole network unfrozen, and we keep our transformer backbone frozen.

Learned Query Embedding. To visualize that our proposed model has indeed learned to disentangle different sound sources through learnable queries, we show t-SNE embeddings of our learnable queries in MUSIC test set [89]. As is shown in Fig. 5, our queries tend to cluster by different instrument classes, learning representative prototypes.

4.3. Extendable Audio Prompt Fine-tuning

Table. 4 evaluates our approach's generalization ability compared with previous methods. We conduct fine-tuning experiments by leave-one-out cross-validation. Baseline models are fine-tuned on the new instrument with all the networks structure unfrozen. With the design of audio prompts discussed in Sec. 3.4, we keep most of our transformer parameters frozen, only fine-tuning the query embedding layer, which has much fewer parameters (0.048% of the total parameters in Transformer).

Fig. 7 (a) shows our performance with a varying num-



(a) (b)

Figure 7. **Fine-tuning curves** of sound separation. (a) Fine-tuning with different number of unseen instrument classes on *MUSIC*. (b) Fine-tuning with different number of unseen event classes on *AVE*.

ber of new instrument classes for fine-tuning on *MUSIC* dataset. We hold out 1, 2, 4, and 6 instrument classes in the pre-training stage and fine-tune our method on these new classes with only the query embedding layer unfrozen. *MUSIC* dataset contains in total of 11 instruments. Notably, our method still yields good results when the network is only pre-trained on 5 instrument types, even fewer than the unseen classes. Fig. 7 (b) shows our model's fine-tuning performance on *AVE* dataset with a varying number of new event classes for fine-tuning. We follow the experimental setup on *MUSIC*, and hold out 2, 4, 6, 8, and 12 event classes for fine-tuning. This demonstrates our model's adaptivity in general sound separation cases.

4.4. Contrastive Verification

Our learnable query-prototypes network is designed to ensure cross-modality consistency and cross-instrument contrast. We assume these prototypes to draw samples of each particular sound class sample close and push away the different prototypes. The question is whether our network design with "visually-named" query trained in the "Mix-and and-Separate" can already achieve this goal? As an alternative, we design an auxiliary contrastive loss for verification: to maximize the cosine similarity of separated audio embedding $\varepsilon_{A_k} = \varepsilon_A \odot M_k$ and the corresponding query embedding Q_k in Q, while minimizing the cosine similarity of separated audio embedding and other query embeddings Q_n (where $n \in [1, N], n \neq k$). We optimize the cross-entropy losses of the cosine similarity scores to obtain contrastive loss $L_{contras}$. To ensure the qualities of audio embedding ε_A and predicted mask M_i are accurate enough, we use a hierarchical task learning strategy [44] to control weights for L_{sep} and $L_{contras}$ at each epoch. The verification loss L_{verify} is: $L_{verify} = w_{sep}(e) \cdot L_{sep} + w_{contras}(e) \cdot L_{contras}$ where e denotes training epoch and $w(e)$ denotes loss weight.

Ablations of auxiliary contrastive loss, shown in Table. 5, demonstrates that our existing design achieves better results without using explicit contrastive loss. This answers the question we raised, that our "visually-named" queries are already contrastive enough for sound disentanglement.

Architecture	SDR↑	SIR↑	SAR↑
w/o lrn.	10.05	14.27	13.71
w/o adpt.	10.89	15.51	14.14
w/ con. best	11.02	15.91	14.10
Ours (w/o con)	11.17	15.84	14.27

Table 5. **Ablations on the auxiliary contrastive loss** on *MUSIC* dataset. "w/o lrn." denotes without learnable linear layer added to queries produced by Transformer decoder; "w/o adpt." denotes that we use a fixed weight for auxiliary contrastive loss without the Hierarchical Task Learning strategy; "w/ con. best" denotes our best model design using auxiliary contrastive loss.

Architecture	SDR↑	SIR↑	SAR↑
Random	6.58	10.79	12.77
Self-audio	10.54	14.81	14.23
Self-motion-audio	10.65	15.37	13.96
Dual-stream	10.46	15.25	13.79
Motion-self-audio	11.17	15.84	14.27

Table 6. **Ablations on the design of Transformer decoder.**

4.5. Ablations of Transformer decoder design

Ablation results of Transformer decoder design on *MUSIC* dataset is shown in Table. 6. "Random" denotes randomly assigning object features to queries, its poor separation result verifies the importance of our "visually-named" queries. "Self-audio" means removing the motion cross attention layer, which confirms the effectiveness of adding the motion feature. We tried two baseline designs against our final selection "Motion-self-audio", as stated in Sec. 3.3. "Self-motion-audio" is a design that puts self-, motion cross-, and audio cross-attention in a single decoder layer. "Dual-stream" means we conduct motion and audio cross-attention in parallel then fuse in the decoder layer. Specific details are in the Supplemental material.

5. Conclusion

We proposed an audio-visual separation method using an adaptable query-based audio mask transformer network. Our network disentangles different sound sources explicitly through learnable audio prototypes initiated by "visually naming". We demonstrate cross-modal consistency and cross-instrument contrast via a multi-modal cross-attention mechanism. When generalizing to new unseen classes, our method can be adapted by inserting additional queries as audio prompts while freezing the attention mechanism. Experiments on both musical and general sound datasets demonstrate performance gain by our iQuery.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018. 2

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. *arXiv preprint arXiv:1907.04975*, 2019. 2

[3] 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 (ECCV)*, pages 208–224, 2020. 1

[4] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 2

[5] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 2

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021. 3

[7] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022. 2

[8] Estefanía Cano, Derry FitzGerald, and Karlheinz Brandenburg. Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pages 1758–1762. IEEE, 2016. 7

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020. 3, 4

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017. 4

[11] Pritish Chandna, Marius Miron, Jordi Janer, and Emilia Gómez. Monoaural audio source separation using deep convolutional neural networks. In *International conference on latent variable analysis and signal separation*, pages 258–266. Springer, 2017. 2

[12] Moitreya Chatterjee, Narendra Ahuja, and Anoop Cherian. Learning audio-visual dynamics using scene graphs for audio source separation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[13] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1204–1213, 2021. 2

[14] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16867–16876, 2021. 2

[15] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 1, 3

[16] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. 1, 3, 4, 6

[17] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3884–3892, 2020. 1

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[19] Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9716–9726, 2022. 4, 6

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[21] 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. 2, 3

[22] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis. *Neural computation*, 21(3):793–830, 2009. 2

[23] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10478–10487, 2020. 1, 3, 4, 5, 7

[24] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. In *European Conference on Computer Vision (ECCV)*, pages 35–53, 2018. 2

[25] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 324–333, 2019. 2

[26] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3879–3888, 2019. 1, 2, 4, 5, 6, 7

[27] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE, 2021. 1, 2, 3

[28] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 2

[29] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2017. 5

[30] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984. 4

[31] Simon Haykin and Zhe Chen. The cocktail party problem. *Neural computation*, 17(9):1875–1902, 2005. 2

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

[33] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9248–9257, 2019. 1

[34] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2136–2147, 2015. 1

[35] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, pages 709–727. Springer, 2022. 2

[36] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 88–95. IEEE, 2005. 2

[37] Kevin Kilgour, Beat Gfeller, Qingqing Huang, Aren Jansen, Scott Wisdom, and Marco Tagliasacchi. Text-driven separation of arbitrary sounds. *arXiv preprint arXiv:2204.05738*, 2022. 2

[38] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 6

[39] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1336–1345, 2021. 2, 3

[40] Jie Hwan Lee, Hyeong-Seok Choi, and Kyogu Lee. Audio query-based music source separation. *arXiv preprint arXiv:1908.06593*, 2019. 2

[41] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[42] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3202–3211, 2022. 3

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[44] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3111–3121, 2021. 8

[45] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 275–285, 2021. 2

[46] Sagnik Majumder and Kristen Grauman. Active audio-visual separation of dynamic sound sources. In *European Conference on Computer Vision (ECCV)*, pages 551–569. Springer, 2022. 2

[47] Mingyuan Mao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, Errui Ding, Baochang Zhang, Shumin Han, et al. Dual-stream network for visual recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:25346–25358, 2021. 3

[48] Josh H McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009. 2

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

[50] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2906–2917, 2021. 3

[51] Juan F Montesinos, Venkatesh S Kadandale, and Gloria Haro. Vovit: Low latency graph-based audio-visual voice separation transformer. In *European Conference on Computer Vision (ECCV)*, pages 310–326. Springer, 2022. 2

[52] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. 2

[53] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1, 2, 5, 7

[54] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition (CVPR), pages 2405–2413, 2016. 2

[55] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *European Conference on Computer Vision (ECCV)*, pages 292–308. Springer, 2020. 2

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 6

[57] 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. 5, 7

[58] Tanzila Rahman, Mengyu Yang, and Leonid Sigal. Tribert: Full-body human-centric audio-visual representation learning for visual sound separation. *arXiv preprint arXiv:2110.13412*, 2021. 1, 3, 4, 5, 7

[59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[60] Sam Roweis. One microphone source separation. *Advances in Neural Information Processing Systems (NeurIPS)*, 13, 2000. 1

[61] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018. 2

[62] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[63] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3222–3231, 2022. 2

[64] Martin Spiertz and Volker Gnann. Source-filter based clustering for monaural blind source separation. In *Proceedings of the 12th International Conference on Digital Audio Effects*, volume 4, 2009. 5, 7

[65] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7262–7272, 2021. 1, 3

[66] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2745–2754, 2021. 2, 3, 4, 5, 6, 7

[67] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision (ECCV)*, pages 436–454. Springer, 2020. 2

[68] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *European Conference on Computer Vision (ECCV)*, pages 247–263, 2018. 2, 5, 6

[69] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, pages 10347–10357. PMLR, 2021. 3

[70] Thanh-Dat Truong, Chi Nhan Duong, Hoang Anh Pham, Bhiksha Raj, Ngan Le, Khoa Luu, et al. The right to talk: An audio-visual transformer approach. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1105–1114, 2021. 2

[71] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Dan 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)*, 2020. 2, 3

[72] Efthymios Tzinis, Scott Wisdom, Tal Remez, and John R Hershey. Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation. In *European Conference on Computer Vision (ECCV)*, pages 368–385. Springer, 2022. 2, 3

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3, 4

[74] 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

[75] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022. 2

[76] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6292–6300, 2019. 2

[77] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19989–19998, 2022. 2

[78] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021. 1, 3

[79] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *IEEE/CVF In-*

ternational Conference on Computer Vision (ICCV), pages 882–891, 2019. 2, 5, 7

[80] Xudong Xu, Hang Zhou, Ziwei Liu, Bo Dai, Xiaogang Wang, and Dahua Lin. Visually informed binaural audio generation without binaural audios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15485–15494, 2021. 2

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

[82] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 558–567, 2021. 3

[83] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16375–16387, 2022. 2

[84] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[85] Renrui Zhang, Han Qiu, Tai Wang, Xuanzhuo Xu, Ziyu Guo, Yu Qiao, Peng Gao, and Hongsheng Li. Monodetr: Depth-aware transformer for monocular 3d object detection. *arXiv preprint arXiv:2203.13310*, 2022. 3

[86] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022. 2

[87] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016. 5

[88] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1735–1744, 2019. 1, 3, 4, 5, 6, 7

[89] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *European Conference on Computer Vision (ECCV)*, pages 570–586, 2018. 1, 2, 3, 4, 5, 6, 7

[90] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16259–16268, 2021. 3

[91] Minghang Zheng, Peng Gao, Renrui Zhang, Kunchang Li, Xiaogang Wang, Hongsheng Li, and Hao Dong. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*, 2020. 3

[92] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6881–6890, 2021. 3

[93] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 2

[94] Dongzhan Zhou, Xinchi Zhou, Di Hu, Hang Zhou, Lei Bai, Ziwei Liu, and Wanli Ouyang. Sepfusion: Finding optimal fusion structures for visual sound separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3544–3552, 2022. 3

[95] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8436–8444, 2021. 2

[96] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision (ECCV)*, pages 350–368. Springer, 2022. 6

[97] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3550–3558, 2018. 2

[98] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation using cascaded opponent filter network. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 5, 7

[99] Lingyu Zhu and Esa Rahtu. Visually guided sound source separation and localization using self-supervised motion representations. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1289–1299, 2022. 1, 3, 4, 5, 6, 7

[100] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3