AudioScopeV2: Audio-Visual Attention Architectures for Calibrated Open-Domain On-Screen Sound Separation (Supplementary Material)

Efthymios Tzinis[1,2⋆], Scott Wisdom[1], Tal Remez[1], and John R. Hershey[1]

[1] Google Research
[2] University of Illinois Urbana-Champaign
etzinis2@illinois.edu, {scottwisdom,johnhershey}@google.com

Outline of the supplementary material

- Section 1: video demos
- Section 2: analysis of calibration
- Section 3: ablations
- Section 4: evaluation on restricted-domain datasets
- Section 5: visualizations of attention maps

1 Video demos

Please see our project website[3] for video demos using our proposed AudioScopeV2 and AudioScope [8]. For these demos, we run the following models trained unsupervised on our proposed unfiltered YFCC100M [7] data: joint and separable SA at 16 FPS, joint and separable CMA at 16 FPS, and AudioScope* (improved version of AudioScope, as described in the main paper) at 16 and 1 FPS. We also use an AudioScope model [8] at 1 FPS, which was trained on filtered YFCC100M. We provide demos on two types of examples: synthetic mixtures of mixtures (MoMs) from the unfiltered random background test set, and single real videos drawn from the test split of unfiltered YFCC100M.

The demos highlight a variety of interesting cases.

- Real Example 1 shows a close-up of a child talking on-screen with strong non-stationary noise in the background; the separable SA models do a remarkable job of suppressing the background noise.
- In real Example 2, there is a child and an adult presumably talking on screen, in the midst of loud off-screen speech from a news broadcast; the separable SA models successfully suppress the off-screen voice while preserving the presumed on-screen subjects of the video, despite their faces being obscured, whereas the AudioScope models yield inconsistent results. Taken together

⋆ Work done during an internship at Google Research.
[3] google-research.github.io/sound-separation/papers/audioscope-v2

these examples illustrate that the models can selectively preserve or block speech depending on whether the talkers are on-screen, even without a frontal view of the on-screen faces; presumably the model is able to use context and other cues to infer which voices correspond to what appears on-screen.
- In real Example 3, a child is playing in a pile of dried leaves with ambient noise in the background.
- In real Example 4, a rocking horse ridden by a child is impacting the walls in a hallway, with some clicking and ambient background noise. In both of these examples, the models are able to differentially suppress the background noise and preserve the on-screen noises, even though both are noise-like sounds.
- The other real examples show operation in low-light conditions (Example 5), selective enhancement of a baseball hit while suppressing background voices (Example 6), and selective enhancement of non-speech eating sounds (Example 7).

In the synthetic demos, Examples 2 and 6 show the model selectively removing or preserving stationary noise depending on the visual input. Synthetic Example 2 shows a person talking on-screen, with off-screen mechanical noise; the models substantially reduce the mechanical noise. Synthetic Example 6 has on-screen rain with off-screen wind noise and voices.

Overall the models show a variety of interesting emergent behaviors, such as preserving inferred on-screen sources, and future work will require a more systematic analysis of conditions amenable to good performance.

2 Analysis of calibration

In this section, we analyze the effect of calibration on the YFCC100M unfiltered test data. We were motivated by the "no processing" baseline providing some counter-intuitive results. As a reminder, we reported "no processing" baselines in Table 1 of our main paper at 0 dB and 6 dB OSR. The 0 dB OSR model just outputs the input audio x as the on-screen estimate $\hat{x}^{\text{on}}$. To calibrate the "no processing" model to 6 dB OSR, we simply use one half the input audio x as the on-screen estimate $\hat{x}^{\text{on}}$.

Counter-intuitively, the 0 dB OSR model achieves a lower median SNR than the 6 dB OSR model. To understand this effect, the left panel of Figure 1 shows a scatter plot of individual examples for 0 dB OSR versus 6 dB OSR. We can see that the median for 6 dB OSR is increased, at the expense of limiting the maximum attainable SNR to 6dB, and results in a drop in the mean SNR.

We also provide a similar plot for the joint CMA AudioScopeV2's configuration, in the right panel of Figure 1, which plots SNR for the uncalibrated model, which achieves 8.7 dB OSR, versus SNR for the model calibrated at 6 dB OSR, following the procedure in Section 4.4 in the main paper. Note that in this case, in contrast to the "no processing" plot, the calibration is adjusting from higher OSR to lower OSR. Unlike the "no processing" case, the calibration procedure on AudioScopeV2 does not cause a drop in mean SNR, and improves SNR performance for examples that had greater than 0dB SNR in the original model. SNR is decreased for

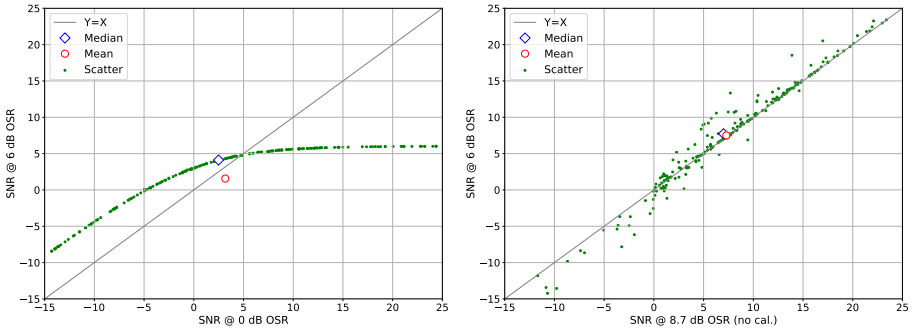

Fig. 1: Left: scatter plot of "no processing" baseline SNRs at 0 dB OSR versus SNRs at 6 dB OSR. Right: scatter plot of the joint CMA model's SNRs before calibration (8.7 dB OSR) versus SNR at 6 dB OSR.

examples that had less than 0dB SNR for the uncalibrated model, but the SNR-limited behavior of the calibrated "no processing" model is avoided.

3 Ablations

This section presents several ablations that we could not include in the main paper due to space constraints.

3.1 Scale-invariant SNR (SI-SNR)

For measuring the fidelity of the reconstruction of the on-screen component, we also report the scale-invariant signal-to-noise ratio (SI-SNR) [6] of the on-screen estimate $\hat{x}^{\text{on}}$, which is defined as follows:

$$\text{SI-SNR}(x, \hat{x}) = 10 \log_{10} \frac{\|\alpha x\|^2}{\|\alpha x - \hat{x}\|^2}, \quad \alpha = \text{argmin}_a \|ax - \hat{x}\|^2 = \frac{x^T \hat{x}}{\|x\|^2}. \tag{1}$$

In the course of our experiments, we discovered that SI-SNR is a potentially optimistic measure of on-screen separation quality, especially when comparing SI-SNR to OSR. Since SI-SNR scales the reference signal to compensate for gain errors on the estimate, this means that a model can predict scaled-down probabilities that maximize OSR without affecting SI-SNR and producing an on-screen estimate that has a gain error. However, SI-SNR is useful in cases where one can obtain higher OSR values by scaling down the estimate and obtain falsely higher OSR numbers (see Table 1).

From Table 1, we see that by halving the gain of the input mixture itself, we are able to increase the SNR 2.5 dB $\rightarrow$ 4.1 dB, which is also close to the performance that the previous state-of-the-art model obtains when trained with our recipe on the unfiltered YFCC100M data (SNR of 5.2 dB). This is due to

Table 1: Evaluation results, including median SI-SNR and SNR, along with the AUC-ROC, for filtered off-screen background (from [8]) and unfiltered random background (our new proposed) test sets at 16 FPS. For each model, calibration to 6 dB OSR is performed separately on the filtered and unfiltered test sets. "AudioScope*" is our improved implementation of AudioScope [8] and using our proposed training procedure while also pre-training the audio source separation module.

	Filtered [8]			Unfiltered (new proposed)		
AV alignment	**SI-SNR**	**SNR**	**AUC**	**SI-SNR**	**SNR**	**AUC**
No processing ($\hat{x}^{\text{on}} = x$ with 0dB OSR)	4.4	4.4	–	2.5	2.5	–
No processing ($\hat{x}^{\text{on}} = x/2$ with 6dB OSR)	4.4	4.7	–	2.5	4.1	–
AudioScope* (previous state-of-the-art)	9.5	5.9	0.77	5.5	5.2	0.71
AudioScopeV2 with Joint CMA (Ours)	**10.8**	**10.0**	**0.85**	**7.8**	**7.7**	**0.84**

SNR providing an overly optimistic measure when the estimate's gain is less than the correct gain [6], and this is the main reason that we also include the SI-SNR metric to all the reported ablation studies.

On the contrary, reporting SI-SNR only could also sometimes lead to false conclusions. For instance, our highest performing model evaluated on the filtered YFCC100M data, joint CMA, clearly outperforms the previous state-of-the-art model by 4.1 (5.9 dB $\rightarrow$ 10.0 dB) in terms of SNR but only by 1.3 (9.5 dB $\rightarrow$ 10.8 dB) in terms of SI-SNR. This is due to the fact that a model is able to always perform higher OSR by scaling down all the estimated probabilities, at the expense of decreasing the gain on the on-screen estimate. As a result, the SNR metric shows that our method is able to estimate a more accurate gain of the on-screen estimate and truly increase its reconstruction fidelity.

3.2 Ablations for proposed separable SA

The hyperparameters for our architecture were chosen according to informal tuning during model development. To determine if there are better settings of our proposed attention-based architectures, we performed a number of ablations. Since results were fairly similar among different architectures, we selected our proposed unsupervised separable SA model as a representative, and retrained this model with a number of different settings. The results are shown in Table 2.

Though there are some settings that work a bit better for separable SA in these ablations (e.g. using fewer attention heads), we decided to present results with the original default settings in our main paper. Generally these results indicate that some additional performance may be available with additional hyperparameter exploration. We do observe that a deeper model with 8 blocks seems to have a negative effect on the performance across the board, perhaps because of overfitting to the training set, or failure to converge.

Table 2: Ablation results, including median SI-SNR, SNR, and OSR, along with the AUC-ROC, for the proposed 16 FPS unsupervised separable SA with 4 attention heads, embedding dimension $D = 128$, $L = 4$ blocks, 8×8 spatial locations from the visual embedding network, and a dropout rate of 0.2. We perform the ablations on the introduced unfiltered random background YFCC100M test partition at 2 calibration points with a specified OSR at 6 and 10 dB. The separation performance of the oracle MixIT* assignment is 10.4 dB SNR and 10.1 dB SI-SNR.

	$\text{OSR}_{\text{target}} = 6$dB				$\text{OSR}_{\text{target}} = 10$dB			
Ablation	**SI-SNR**	**SNR**	**OSR**	**AUC**	**SI-SNR**	**SNR**	**OSR**	**AUC**
–	7.3	6.6	6.0	0.80	6.6	5.5	10.0	0.80
4x4 spatial	8.4	7.0	6.0	0.83	7.8	5.8	10.0	0.83
No dropout	6.9	7.0	6.0	0.81	7.0	6.1	10.0	0.81
2 heads	8.1	7.4	6.0	0.83	8.1	6.3	10.0	0.83
8 heads	7.2	6.2	6.0	0.77	6.7	4.7	10.0	0.77
2 blocks	7.1	6.5	6.0	0.80	7.2	5.3	10.0	0.80
8 blocks	2.5	3.5	1.1	0.48	2.5	3.5	1.1	0.48
D=256	7.6	6.8	6.0	0.78	8.2	5.7	10.0	0.78
D=64	7.4	7.2	6.0	0.80	7.5	6.0	10.0	0.80

3.3 Unsupervised vs semi-supervised results

We also experimented with semi-supervised training, where we leverage videos annotated for the presence of on-screen and off-screen sounds. Semi-supervised training uses unsupervised NOn examples as described in the main paper, plus additional "human-labeled on-screen-only (LOn)" and "human-labeled off-screen-only (LOff)" examples. For these examples we use both single-mixture and MoM versions. LOn single-mixture examples are just video frames and audio drawn from a unanimously-rated on-screen-only video. LOn MoM examples are the same as LOn single-mixture, except that synthetic off-screen audio from a random video is added. LOff single-mixture and MoM examples are the same as LOn examples, except that unanimously-rated off-screen-only videos are used for the primary video frames and audio. An exact cross-entropy loss is used for training the on-screen classifier with these labeled examples, where the MixIT assignments are used as classifier labels y for on-screen examples, and the classifier labels y are set to zero for off-screen examples. For semi-supervised training, examples in the batch are dynamically sampled, with 50% NOn MoM, 12.5% LOn single-mixture, 12.5% LOn MoM, 12.5% LOff single-mixture, and 12.5% LOff MoM.

We show the results using our proposed models at 16 FPS and at 2 different OSR target levels in Table 3. Our results are somewhat mixed. For separable models, AUC-ROC improves by up to 0.09, but does not improve for joint models. Also, even when AUC-ROC improves, this does generally improve calibrated on-screen reconstruction in terms of SI-SNR and SNR. We think this is due to the fact that AUC-ROC is measured across all the possible operating points of

Table 3: Evaluation results, including median SI-SNR and SNR, along with the AUC-ROC, on the unfiltered random background test set at 16 FPS for the proposed models for unsupervised or semi-supervised training. Models are calibrated to 2 different OSR levels: 6 dB and 10 dB. The detection capability of the models for the on-screen presence of the estimated sources, measured by the weighted area under the ROC curve (AUC), remains unaltered with different specified OSR operating points.

		Semi-	$\text{OSR}_{\text{target}} = 6\text{dB}$		$\text{OSR}_{\text{target}} = 10\text{dB}$		
AV alignment		Supervised	**SI-SNR**	**SNR**	**SI-SNR**	**SNR**	**AUC**
SA	Joint		7.7	7.7	7.6	6.9	0.83
		✓	6.5	5.0	6.5	5.5	0.80
	Sep.		7.3	6.6	6.6	5.5	0.80
		✓	7.2	7.0	8.0	6.4	0.88
CMA	Joint		**7.8**	7.7	**8.3**	6.7	0.84
		✓	6.2	6.5	6.4	5.6	0.83
	Sep.		7.5	7.1	7.2	5.8	0.80
		✓	6.3	6.5	7.4	6.7	**0.89**

the classifier, whereas the SNR evaluation can only be obtained after choosing a specified OSR target level. We postulate that one might be able to extend our work using appropriate signal-level reconstruction losses and also improve reconstruction fidelity performance for semi-supervised cases, but we defer this to future work.

3.4 Ablation on selected calibration points

Another important aspect of our proposed calibration method is that the tolerance of off-screen sound interference can be specified. To that end, we show the importance of our method by comparing to uncalibrated model evaluations at a few interesting OSR operating points. The results of this ablation study are in Table 4.

Notice that the uncalibrated joint SA seems to be performing significantly better than separable SA in terms of on-screen SNR: 7.1 dB vs 5.3 dB. However, if we also consider the ability of these models to suppress the off-screen component, the comparison becomes less clear since joint SA and separable SA obtain 8.3 dB and 10.8 dB OSR, respectively. To allow a more fair and easy-to-understand comparison of these models, our proposed calibration method can compensate for the OSR mismatch by tuning both models to specified OSR target level (e.g. $\text{OSR}_{\text{target}} = 6\text{dB}$). After doing this, we can see that the actual SNR difference is almost 1 dB (joint SA and separable SA obtain 7.7 dB and 6.6 dB SNR, respectively).

Unsurprisingly, for increasing levels of specified OSR target levels, the SNR performance on the reconstruction of the on-screen component gradually declines.

Table 4: Evaluation results, including median SI-SNR, SNR, and OSR, along with the AUC-ROC, on the unfiltered random background test set at 16 FPS for the proposed models with different levels of calibrated OSR levels. All models have been trained using our proposed unsupervised learning procedure on the unfiltered dataset YFCC100M train partition while also pre-training the audio source separation network. We show the results when we use the raw pre-trained models with no calibration and when we calibrate AudioScopeV2 at 3 different OSR levels of 6, 10, and 15 dB. The detection capability of the models for the on-screen presence of the estimated sources, measured by the weighted area under the ROC curve (AUC), remains unaltered with different specified OSR operating points.

Calibration	AV alignment		**SI-SNR**	**SNR**	**OSR**	**AUC**
Uncalibrated	SA	Joint	7.8	7.1	8.3	0.83
		Sep.	6.7	5.3	10.8	0.80
	CMA	Joint	**8.3**	7.0	8.7	**0.84**
		Sep.	7.2	5.8	9.3	0.80
$\text{OSR}_{\text{target}} = 6\text{dB}$	SA	Joint	7.7	**7.7**	6.0	0.83
		Sep.	7.3	6.6		0.80
	CMA	Joint	7.8	**7.7**		**0.84**
		Sep.	7.5	7.1		0.80
$\text{OSR}_{\text{target}} = 10\text{dB}$	SA	Joint	7.6	6.9	10.0	0.83
		Sep.	6.6	5.5		0.80
	CMA	Joint	**8.3**	6.7		**0.84**
		Sep.	7.2	5.8		0.80
$\text{OSR}_{\text{target}} = 15\text{dB}$	SA	Joint	7.6	6.0	15.0	0.83
		Sep.	7.0	3.4		0.80
	CMA	Joint	**8.3**	5.5		**0.84**
		Sep.	7.2	4.3		0.80

However, we want to emphasize that in practice the operating point can be calibrated on validation data, according to the needs of an application, or according to user preferences. We also see that the separable versions of SA and CMA are able to perform on-par with the much more computationally expensive joint counterparts across all the specified OSR target levels.

4 Evaluation on restricted-domain datasets

Though our goal is to train a general-purpose on-screen separation model, it is interesting to see how well AudioScopeV2 performs on specialized domains without any further fine-tuning. We evaluated our models on these datasets, but were unable to include these results in the main paper due to space constraints. In the following subsections, we present these results.

Our purpose is not to compete with prior methods that use carefully curated training data to match these test sets (e.g. Mandarin has a very small corresponding training set, and approaches in the literature have crafted custom training sets) as we anticipate that our general model will do less well than a model specifically trained towards a more specialized task and domain. We show these mismatched evaluations to examine whether more general approaches like ours could be used in handling more specialized tasks.

4.1 MUSIC dataset for audio-visual musical instrument separation

To measure performance on a non-speech task, we evaluated our proposed models on the MUSIC dataset [12], which is a dataset of single-source videos of people playing musical instruments. We used the standard protocol [3] to prepare the dataset, where we created 10 mixtures for each of the 55 possible pairs of 11 instrument classes, for a total of 550 examples. For each example, we use the video for one of the instruments as the video input to AudioScopeV2, and we do this for both videos (thus the total number of examples is $2 \cdot 550 = 1100$). Performance is measuring using `bss_eval_sources` [9], which measures signal-to-distortion ratio (SDR), signal-to-inference ratio (SIR), and signal-to-artifact ratio (SAR). These measures find an optimal time-invariant 512-tap filter that can be applied to the reference to maximize SDR. We compare our methods to a number of other recent approaches that also evaluate on this dataset.

There are a few things to notice from these results. First, neither our proposed methods nor AudioScope reach state-of-the-art performance on this task compared to models trained on matched data with various degrees of supervision. However, the oracle MixIT* performance of the audio-only separation component of our proposed models, where separated sources are assigned to one of the ground-truth reference audio sources, is quite strong, within only 1.4 dB SDR for the best matched-training model (10.0 dB versus 11.4 dB). This MixIT* performance is also better than AudioScope, which only achieves 8.8 dB SDR. This improvement over AudioScope is presumably due to audio-only pre-training of our separation model on unfiltered YFCC100M with MixIT.

Unfortunately, the non-oracle $\hat{x}^{\text{on}}$ output, which uses the predicted on-screen probabilities as mixing weights, still lags behind the oracle MixIT* scores. However, our proposed $\hat{x}^{\text{on}}$ models do achieve better performance than AudioScope: in the best case, unsupervised joint CMA achieves 3.1 dB SDR, compared to -0.5 dB SDR for AudioScope, which is a significant boost. This suggests that using a wider variety of YFCC100M data helps AudioScopeV2 generalize, but that there may still be mismatch between videos in the MUSIC dataset and videos in YFCC100M. We anticipate that fine-tuning one of our proposed models on data from a target domain could help reduce this gap in performance.

Table 5: Results on the MUSIC dataset, including mean SDR, SIR, and SAR.

Model	Oracle?	Training data	SDR	SIR	SAR
Sound-of-Pixels [12]		MUSIC	5.4	11.0	9.8
Sound-of-Motions [11]		MUSIC	4.8	11.0	8.7
MP-Net [10]		MUSIC	5.7	11.4	10.4
Co-Separation [3]		MUSIC	7.4	13.7	10.8
Cascaded Opponent Filter [14]		MUSIC	10.1	16.7	13.0
A(Res-50, att) + S(DV3P) [13]		MUSIC	9.4	15.6	12.7
A(Res-50, class.) + S(DV3P) [13]		MUSIC	10.6	17.2	12.8
AVSGS [1]		MUSIC	**11.4**	**17.3**	13.5
AudioScope [8] $\hat{x}^{\text{on}}$		Filtered YFCC100M	-0.5	2.8	11.2
AudioScope [8] MixIT*	✓	Filtered YFCC100M	8.8	13.0	13.1
Joint SA (unsup), MixIT*	✓	YFCC100M	10.0	14.1	14.6
Joint SA (unsup), $\hat{x}^{\text{on}}$		YFCC100M	2.1	3.6	18.1
Joint SA (semi-sup), MixIT*	✓	YFCC100M	9.8	13.8	14.7
Joint SA (semi-sup), $\hat{x}^{\text{on}}$		YFCC100M	1.7	3.3	18.3
Sep. SA (unsup), MixIT*	✓	YFCC100M	10.0	14.1	14.4
Sep. SA (unsup), $\hat{x}^{\text{on}}$		YFCC100M	0.4	1.3	19.7
Sep. SA (semi-sup), MixIT*	✓	YFCC100M	9.8	13.8	14.6
Sep. SA (semi-sup), $\hat{x}^{\text{on}}$		YFCC100M	0.4	1.6	18.1
Joint CMA (unsup), MixIT*	✓	YFCC100M	9.7	13.9	14.6
Joint CMA (unsup), $\hat{x}^{\text{on}}$		YFCC100M	3.1	4.8	18.6
Joint CMA (semi-sup), MixIT*	✓	YFCC100M	9.8	13.8	14.5
Joint CMA (semi-sup), $\hat{x}^{\text{on}}$		YFCC100M	2.1	3.5	19.1
Sep. CMA (unsup), MixIT*	✓	YFCC100M	9.9	13.9	14.5
Sep. CMA (unsup), $\hat{x}^{\text{on}}$		YFCC100M	1.9	3.3	17.5
Sep. CMA (semi-sup), MixIT*	✓	YFCC100M	9.9	14.0	14.4
Sep. CMA (semi-sup), $\hat{x}^{\text{on}}$		YFCC100M	0.0	0.3	**25.4**

4.2 Mandarin dataset for audio-visual speech enhancement

Audio-visual speech enhancement, which is the task of separating speech of an on-screen talker given video and noisy speech audio, has been explored by many recent works. To measure performance on this more restricted task, we use the Mandarin dataset [5], which consists of video clips of 3-5 seconds of a person speaking Mandarin sentences, where background noise has been artificially added. The results are shown in Table 6, which includes several models from recent works which were trained specifically for the audio-visual speech enhancement task.

Table 6: Results on Mandarin audio-visual speech enhancement evaluation set, in terms of mean SDR.

Model	Oracle?	Training task	SDR
Hou et al. [5]		AV speech enhancement	2.8
Ephrat et al. [2]		AV speech enhancement	6.1
Gao and Grauman [4]		AV speech enhancement	6.7
AudioScope [8] $\hat{x}^{\text{on}}$		AV universal on-screen sep.	2.5
AudioScope [8] MixIT*	✓	AV universal on-screen sep.	3.4
Joint SA (unsup), MixIT*	✓	Audio-only universal on-screen sep.	9.8
Joint SA (unsup), $\hat{x}^{\text{on}}$		AV universal on-screen sep.	1.5
Joint SA (semi-sup), MixIT*	✓	Audio-only universal on-screen sep.	9.8
Joint SA (semi-sup), $\hat{x}^{\text{on}}$		AV universal on-screen sep.	-0.2
Sep. SA (unsup), MixIT*	✓	Audio-only universal on-screen sep.	9.6
Sep. SA (unsup), $\hat{x}^{\text{on}}$		AV universal on-screen sep.	-0.1
Sep. SA (semi-sup), MixIT*	✓	Audio-only universal on-screen sep.	10.0
Sep. SA (semi-sup), $\hat{x}^{\text{on}}$		AV universal on-screen sep.	0.6
Joint CMA (unsup), MixIT*	✓	Audio-only universal on-screen sep.	9.6
Joint CMA (unsup), $\hat{x}^{\text{on}}$		AV universal on-screen sep.	2.3
Joint CMA (semi-sup), MixIT*	✓	Audio-only universal on-screen sep.	**10.0**
Joint CMA (semi-sup), $\hat{x}^{\text{on}}$		AV universal on-screen sep.	-0.5
Sep. CMA (unsup), MixIT*	✓	Audio-only universal on-screen sep.	9.7
Sep. CMA (unsup), $\hat{x}^{\text{on}}$		AV universal on-screen sep.	1.8
Sep. CMA (semi-sup), MixIT*	✓	Audio-only universal on-screen sep.	9.9
Sep. CMA (semi-sup), $\hat{x}^{\text{on}}$		AV universal on-screen sep.	1.1

First, note that our oracle MixIT* models outperform state-of-the-art on this dataset (10.0 dB SDR, versus 6.7 dB SDR for the best matched-training baseline). This implies that the general-purpose separation model pre-trained with MixIT is quite strong. Also, pre-training on unfiltered YFCC100M is perhaps why the MixIT* performance of our models is so much better than the MixIT* performance of AudioScope, which was only trained on filtered YFCC100M.

However, as with the MUSIC dataset, the non-oracle output of $\hat{x}^{\text{on}}$ degrades compared to the oracle performance. We postulate that due to the mismatch

between the Mandarin video data and YFCC100M video data, our proposed models score 2.3 dB in terms of SDR which could be potentially boosted using fine-tuning on matched video samples.

5 Visualizations of attention maps

In Figure 2 several attention maps are displayed for the proposed audio-visual attention architectures. Each heatmap is derived by using the per-source on-screen classifier probability score to weight the attention map for the corresponding input frame while also summing across the different heads. In most of the displayed examples, the warm color regions co-locate with regions of the video frame that represents on-screen objects. Note that for the unsupervised models (two columns on the left), the attention maps are more dispersed across the image and sometimes focused on the background of each image (see rows 2-4 at the last column). We postulate that with strongly labeled data and/or pre-trained segmentation models one could possibly sharpen the accuracy of such attention mechanisms, but we defer that to future work.

Although the analysis of the attention maps for the on-screen objects hints at the capability of the model to understand the audio-visual alignment, we would like to better understand the implicit representation obtained from the SA and CMA models. To help with this, several attention maps obtained by different heads of the first layer of the semi-supervised separable SA model and the semi-supervised separable CMA model are shown in Figures 3 and 4, respectively. Notice that in both cases of attention architectures there are heads which attend to different parts of the input frame potentially showing the expressiveness of the proposed layers even in the case of the more efficient separable variation. Interestingly, “Head 1” (third column) in Figure 3 seems to have a more disperse attention pattern, possibly showing that this head learns to attend more to the background compared to other heads e.g. “Head 0”.

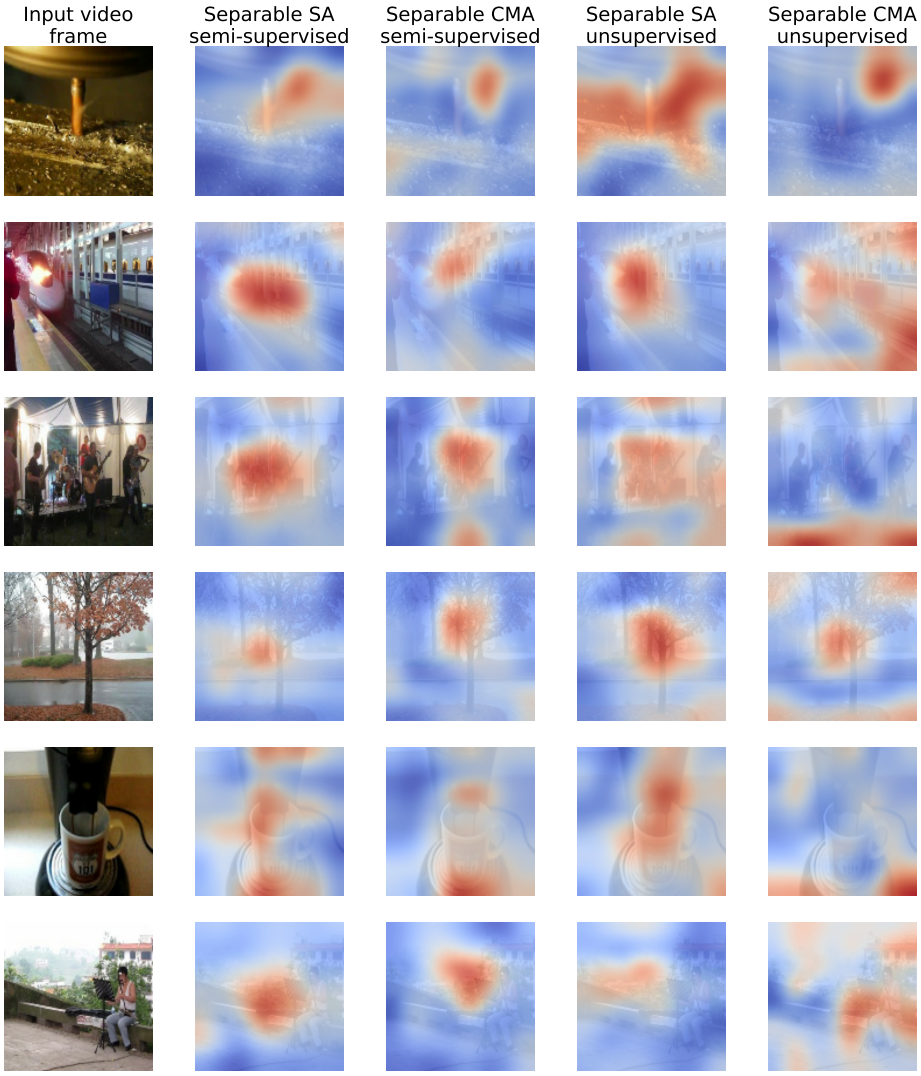

Fig. 2: Weighted attention maps for different training conditions with the proposed audio-visual attention models using uncalibrated separable self-attention (SA) and cross-modal attention (CMA) for a given input video frame. The attention maps have been weighted using the on-screen estimated probability per corresponding source.

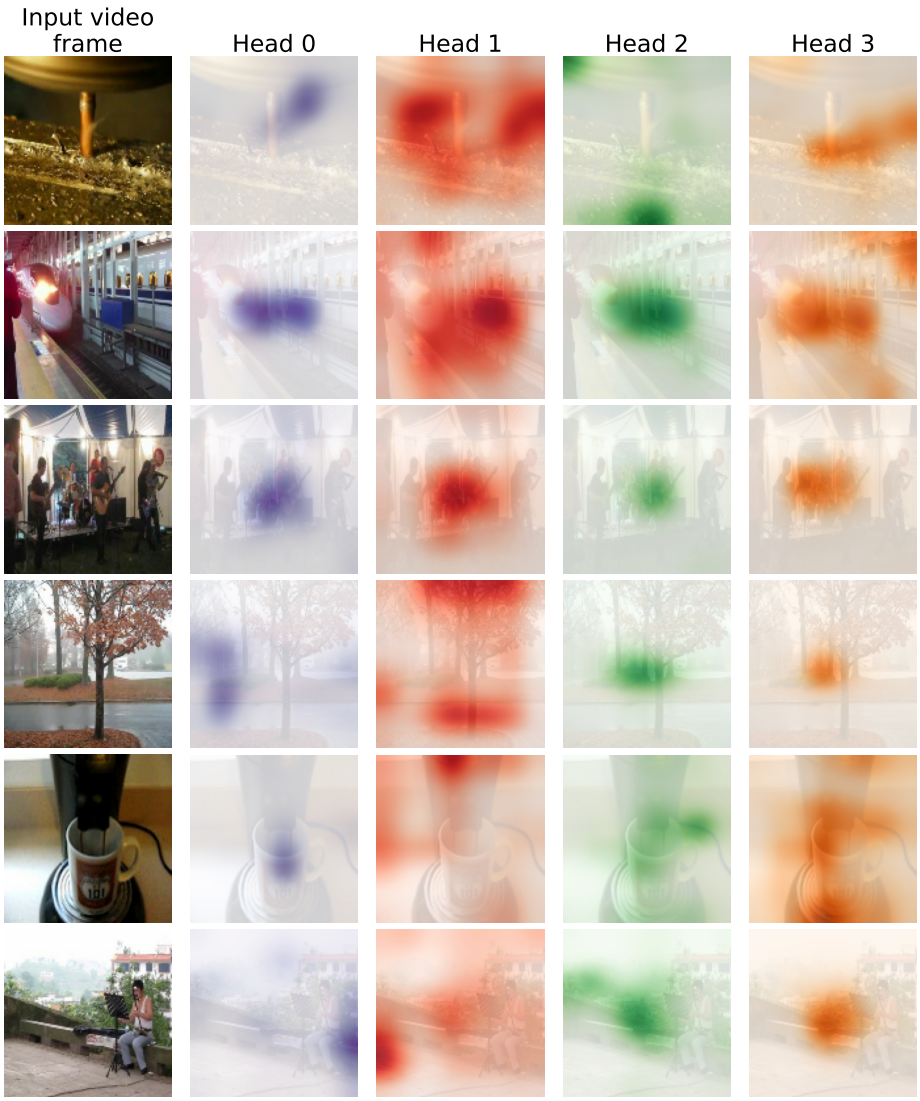

Fig. 3: Attention maps obtained from the different heads of the first separable audio-visual self-attention (SA) layer.

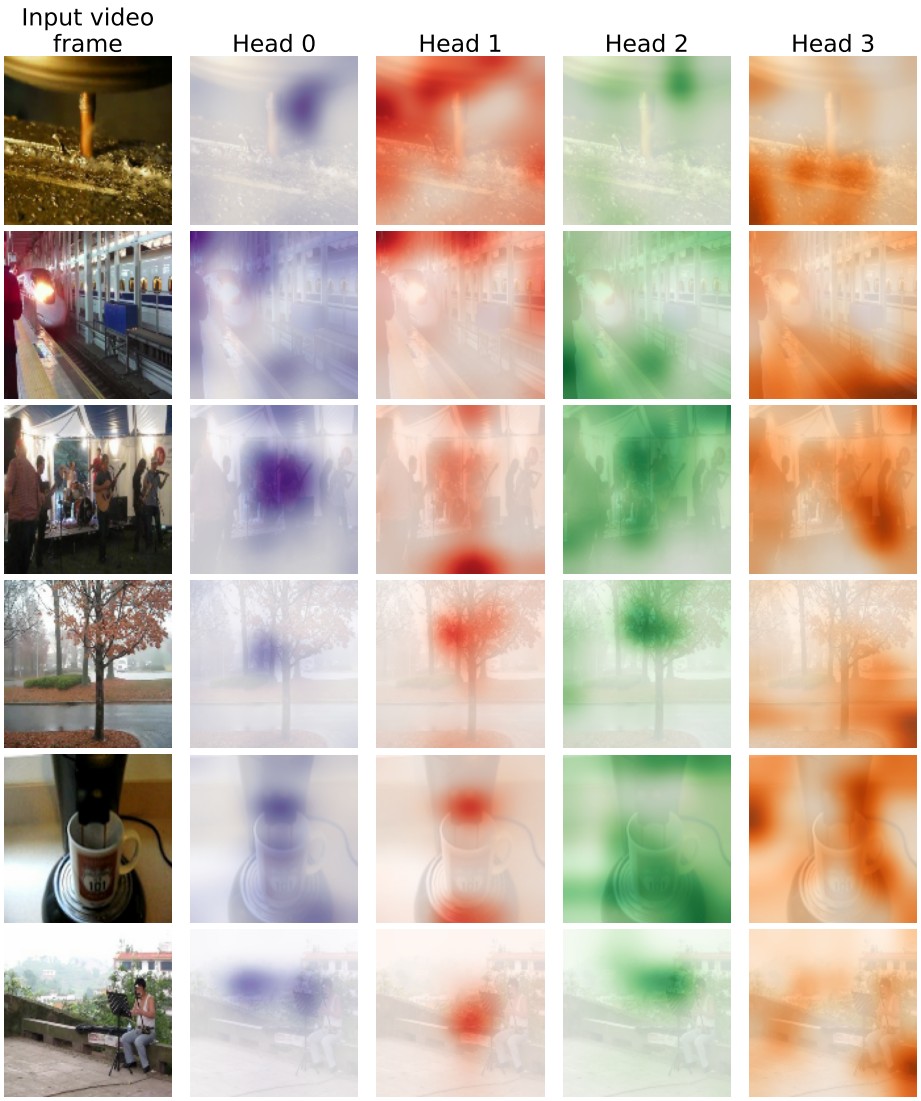

Fig. 4: Attention maps obtained from the different heads of the first separable audio-visual cross-modal attention (CMA) layer.

Attributions

Still images from the following videos are used in Figures 2, 3, and 4:
"Tiny End Mill" by ukweli, license: CC-BY-SA 2.0
"Shinkansen" by pauldesu.com, license: CC-BY 2.0
"A Sample of Broken Mouth Annie" by Tobyotter, license: CC-BY 2.0
"A rainy autumn afternoon in Norcross" by sylvar, license: CC-BY 2.0
"3rd frothy Cup of java, good morning." by miheco, license: CC-BY-SA 2.0
"MVI_6134" by kenner116, license: CC-BY 2.0

References

1. Chatterjee, M., Le Roux, J., Ahuja, N., Cherian, A.: Visual scene graphs for audio source separation. In: Proc. IEEE International Conference on Computer Vision (ICCV). pp. 1204–1213 (2021)
2. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM TOG **37**(4), 1–11 (2018)
3. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3879–3888 (2019)
4. Gao, R., Grauman, K.: VisualVoice: Audio-visual speech separation with cross-modal consistency. In: Proc. IEEE International Conference on Computer Vision (CVPR). pp. 15490–15500 (2021)
5. Hou, J.C., Wang, S.S., Lai, Y.H., Tsao, Y., Chang, H.W., Wang, H.M.: Audio-visual speech enhancement using multimodal deep convolutional neural networks. Proc. IEEE Transactions on Emerging Topics in Computational Intelligence **2**(2), 117–128 (2018)
6. Le Roux, J., Wisdom, S., Erdogan, H., R. Hershey, J.: SDR–half-baked or well done? In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 626–630 (2019)
7. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
8. Tzinis, E., Wisdom, S., Jansen, A., Hershey, S., Remez, T., Ellis, D.P., Hershey, J.R.: Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In: Proc. International Conference on Learning Representations (ICLR) (2021)
9. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **14**(4), 1462–1469 (2006)
10. Xu, X., Dai, B., Lin, D.: Recursive visual sound separation using minus-plus net. In: Proc. IEEE International Conference on Computer Vision (CVPR). pp. 882–891 (2019)
11. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. In: Proc. IEEE International Conference on Computer Vision (CVPR). pp. 1735–1744 (2019)
12. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European conference on computer vision (ECCV). pp. 570–586 (2018)

13. Zhu, L., Rahtu, E.: Separating sounds from a single image. arXiv preprint arXiv:2007.07984 (2020)
14. Zhu, L., Rahtu, E.: Visually guided sound source separation and localization using self-supervised motion representations. In: Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1289–1299 (2022)