ACAV-1M: Data Curation and Benchmarking for Audio-Visual Representation Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

The natural alignment of visual and audio information in videos provides a strong learning signal. However, commonly used large-scale video datasets contain audio-visual signals that are not aligned, e.g. background music. This limits the development of robust models that leverage the complementary nature of audio and video data. To address this limitation, we curate *ACAV-1M*, a new large-scale dataset that contains one million samples sourced from the ACAV-100M dataset. The *ACAV-1M* dataset is obtained through a pipeline that ensures the audio-visual correspondence and synchronization of samples in the dataset. Our pipeline transforms raw video and audio into text captions, followed by text summarization and an extensive filtering procedure. The filtering is done based on audio-caption alignment, audio-visual instance semantic alignment, and temporal synchronization. Furthermore, we propose an audio-visual learning benchmark that supports a diverse range of downstream tasks. Empirical evaluations demonstrate that models trained on *ACAV-1M* achieve superior performance compared to using existing datasets across all tasks. Our *ACAV-1M* dataset and code to reproduce all benchmark results will be made publicly available upon acceptance.

1 Introduction

Recent advancements in multimodal learning, exemplified by the Flan Collection [1] and MMC4 [2], have showcased significant strides for multimodal vision-language models. These developments underline the potential of integrated multimodal datasets for enhancing model performance. However, the field lacks a high-quality, large-scale collection specifically tailored for audio-visual learning, where audio and visual data complement each other to achieve a more holistic understanding of the environment.

The importance of a unified audio-visual dataset stems from the need for a systematic approach to evaluatethe interaction between audio and visual inputs, which are often treated independently, as shown in Table 1. The integration of these modalities promises to improve the robustness and accuracy of learning models by leveraging their inherent complementary properties [3, 4, 5, 6, 7]. The absence of such datasets might hamper the development of audio-video models that can effectively exploit the synergies between sight and sound, thus limiting advancements in this area.

To address this, we introduce a new dataset, namely *ACAV-1M*, which consists of one million audio-visual samples. Our *ACAV-1M* follows a curation pipeline that consists of several steps. First, a multimodal Large Language Model (LLM) [8] generates multiple captions from audio and video

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Table 1: Details about dataset source, modality, number of samples, and benchmark tasks.

Dataset	Modality	# Data	Benchmark Tasks
ACAV100M [11]	Audio, Video	100M	Classification
AudioSet [12]	Audio, Video	2.1M	Classification
Flickr-SoundNet [3]	Audio, Video	2M	Classification, Localization
VGG-Sound [13]	Audio, Video	200K	Classification, Localization
AudioCaps [14]	Audio, Video	48K	Retrieval
Kinetics-Sound [15]	Audio, Video	19K	Classification
LLP [16]	Audio, Video	12K	Video Parsing
AVSD [17]	Audio, Video, Text	12K	Scene-Aware Dialog
MUSIC-AVQA [18]	Audio, Video, Text	9K	Question-Answering
AVS-Bench [19]	Audio, Video	7K	Segmentation
Clotho [20]	Audio, Text	5K	Retrieval
AVE [21]	Audio, Video	4K	Localization
MUSIC [22]	Audio, Video	448	Source Separation
ACAV-1M (ours)	Audio, Video, Text	1M	Cls. & SrcLoc. & Retrieval & SADialog. VideoPars. & QA & Seg. & SrcSep.

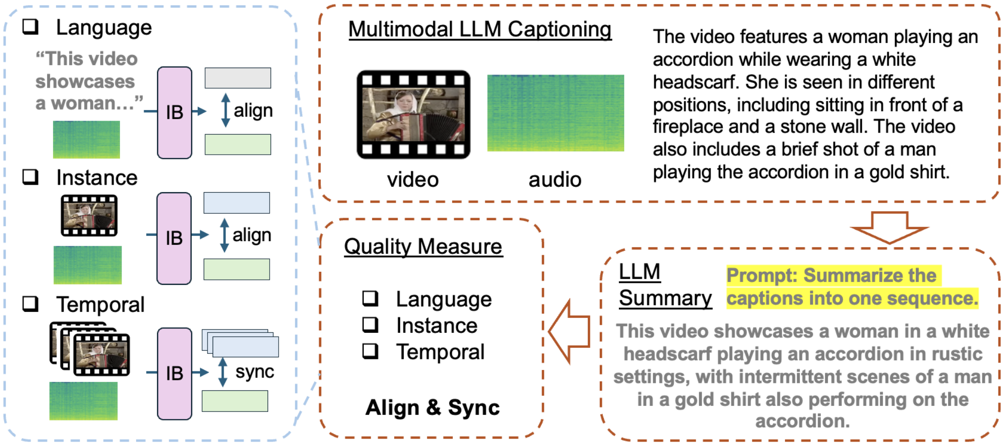
inputs. Then, we use an LLM [9] to summarize the long captions into one sentence for the following quality measure steps. Lastly, we utilize ImageBind [10] to measure audio-language, audio-video instance, and audio-video temporal alignment for data curation. For audio-language alignment, we compute the normalized cosine similarity between audio instance features and caption features extracted from ImageBind [10]. For audio-video instance alignment, we calculate the normalized cosine similarity between audio instance features and video features extracted from ImageBind. For audio-video temporal alignment, we compute the normalized cosine similarity between audio instance features and video features across all ten seconds extracted from ImageBind. These final alignment quality check steps ensure the coherence and synchronization between modalities.

Our dataset and benchmark not only provide tools for measuring data quality on audio-visual instance alignment and temporal alignment, but also support an extensive range of downstream applications. These include audio-visual classification, sound source localization, retrieval, video parsing, scene-aware dialogue, audio-visual question-answering, segmentation, and sound source separation. We provide benchmark results for each of these applications with task-specific methods, and each of these applications is backed by benchmark baselines, task-specific methods, and both pre-trained and novel multimodal foundation models developed using *ACAV-1M*.

Empirical results from extensive experiments demonstrate that models trained on *ACAV-1M* surpass existing methods, highlighting the dataset's effectiveness and scalability properties. This establishes *ACAV-1M* as a significant step towards the systematic integration of audio and visual data in machine learning research, providing a robust platform for exploring new frontiers in multimodal interaction and representation learning.

To summarize, we make the following four contributions:

- We curate the *ACAV-1M* dataset with one million audio-visual samples designed to address the gap in existing multimodal datasets for audio and visual data.
- Our data curation pipeline is a novel contribution that includes the transformation of raw video and audio into detailed, aligned captions using a multimodal large language model.
- We establish comprehensive benchmarks and task-specific methods that leverage our dataset to advance the state-of-the-art in audio-visual learning.
- Extensive experimental analyses demonstrate the effectiveness and scalability of models on *ACAV-1M* compared to existing audio-visual datasets.



The video features a woman playing an accordion while wearing a white headscarf. She is seen in different positions, including sitting in front of a fireplace and a stone wall. The video also includes a brief shot of a man playing the accordion in a gold shirt.

This video showcases a woman in a white headscarf playing an accordion in rustic settings, with intermittent scenes of a man in a gold shirt also performing on the accordion.

Figure 1: Illustration of the proposed ACAV-1M collection paradigm. First, we utilize a multimodal Large Language Model (LLM) [9] to generate multiple captions from audio and video inputs. Then, an LLM is adopted to summarize the long captions into one sentence for the following alignment filtering steps. Finally, we use ImageBind (IB) [10] to ensure the coherence and synchronization between modalities by measuring audio-language, audio-video instance, and audio-video temporal alignment.

2 Related Work

Multimodal benchmarks. Dataset curation efforts, such as the Flan Collection [1] and mmc4 [2], have set precedents for multimodal learning. The Flan Collection has been instrumental in effective instruction tuning, while mmc4 addresses the challenges of few-shot, in-context, and interleaved learning across visual and language models. These benchmarks have laid the groundwork for ACAV-1M emphasizing the need for datasets that support intricate multimodal interactions.

Audio-visual learning. Audio-visual representations learning has been addressed in many previous works [3, 4, 5, 6, 7, 22, 23, 24, 25, 26, 27, 28, 29, 30]. Exploiting the natural alignment across the audio and visual modalities is beneficial for many audio-visual tasks, such as audio-event localization [21, 31, 32, 33], audio-visual localization [34, 35, 36, 25], audio-visual navigation [36, 37, 38], and audio-visual parsing [16, 39, 40, 41]. Different to the aforementioned methods that focus on downstream applications, we propose a new dataset that gives boosts on downstream tasks when used for pre-training.

Audio-visual benchmarks. Existing audio-visual datasets, such as AudioSet [12], VGGSound [13], and ACAV100M [11], provide valuable resources for training and testing audio-visual models. These datasets have advanced audio-visual learning but are limited in size or contain noisy data or labels. Our ACAV-1M complements existing datasetsby offering a structured and aligned dataset that facilitates cleaner multimodal integration.

3 ACAV-1M dataset and benchmark

3.1 Datasets Construction and Statistics

ACAV-1M was meticulously constructed with a dataset curation process that ensures the close alignment between audio and visual elements which is crucial for complex multimodal learning tasks.

Data curation. The dataset curation follows a robust pipeline, starting with raw audio and video data, as illustrated in Figure 1. Each video clip is processed through our multimodal Large Language Model (LLM) to extract descriptive captions. Specifically, VideoLLaVA [8] generates several sentence-level descriptions. These are condensed by a general LLM [9] into a single, comprehensive caption that captures the essence of the audio-visual content. This method ensures that our dataset supports semantic analysis and retrieval tasks effectively.

Data statistics. *ACAV-1M* is annotated with captions for both the audio and video components, facilitating cross-modal training and evaluation. Each audio segment is set to a 10-second duration to standardize the dataset and simplify the processing requirements. Audio data is categorized into various classes such as music, nature sounds, animal sounds, speech, machine noises, and others, providing a broad spectrum of audio types for comprehensive multimodal learning.

Alignment filtering. We quantify the alignment across modalities in our quality measure criteria. We employ ImageBind [10] to ensure several forms of alignment. 1) Language Alignment: The alignment between text captions and both audio and visual content is assessed with a normalized cosine similarity threshold of 0.5, ensuring that descriptions accurately reflect the content. 2) Instance Alignment: The synchronization between audio and visual streams is verified, with an emphasis on maintaining a normalized cosine similarity threshold of 0.5 to ensure alignment across modalities. 3) Temporal Alignment: Audio and visual data are aligned within a temporal window of 1 second per segment, with an average alignment threshold of 0.5 across the dataset.

3.2 Audio-visual benchmark

Audio-visual tasks. *ACAV-1M* supports an extensive range of audio-visual downstream tasks, each designed to leverage the rich, multimodal nature of the *ACAV-1M* dataset.

1. **Audio-Visual Classification.** The goal is to classify the scenes or objects depicted in the audio-visual clips with accuracy as evaluation. For linear probing and fine-tuning on audio-visual classification, we used VGGSound-Music with 49 classes and VGGSound-All with 221 categories.
2. **Audio-Visual Source Localization.** This task measures the model's ability to localize sound sources within a visual frame, assessed by the mean Intersection over Union (mIoU). We use Flickr-SoundNet [7] with 4,500 pairs for training and testing the model on 250 audio-visual pairs of sounding objects and extended 250 non-sounding objects introduced in SLAVC [42].
3. **Audio-Visual Retrieval.** The focus is the recall of relevant audio-visual content based on query descriptions. We use MSR-VTT [43] that includes 10K YouTube videos with 200K description sentences, where 9K is split for training and 1K for testing.
4. **Audio-Visual Scene-Aware Dialog.** This task focuses on generating dialogues that are contextually relevant to given audio-visual scenes, evaluated via BLEU and METEOR scores. We use the AVSD track of the 10-th Dialog System Technology Challenges (DSTC10) [44] dataset.
5. **Audio-Visual Video Parsing.** This involves parsing complex video scenes into simpler segments, evaluated using an F-score at a mIoU threshold of 0.5. The LLP dataset [16] contains 11,849 YouTube video clips of 10-seconds long from 25 different event categories, such as car, music, cheering, speech, etc. We follow the official splits [16] of validation and test sets to train and test.
6. **Audio-Visual Question-Answering.** This task tests accuracy in answering questions based on the content depicted in the audio-visual clips. we use the MUSIC-AVQA [18] dataset that consists of 45,867 question-answer pairs and 9,288 videos.
7. **Audio-Visual Segmentation.** This task focuses on the segmentation masks of visual elements, with performance measured by the F1 Score. AVSBench [19] includes 4,932 videos (in total 10,852 frames) from 23 categories, including instruments, humans, animals, etc. We use the official split of 3,452/740/740 videos for train/val/test.
8. **Audio-Visual Source Separation.** The objective is to measure the ability to isolate individual audio sources from a mixed audio track, evaluated using metrics such as Signal to Distortion Ratio (SDR), Signal to Interference Ratio (SIR), and Signal to Artifacts Ratio (SAR). We use VGGSound-Music [45] with 40,908 video clips from 49 music categories for training and 1201 clips for testing. VGGSound-Instruments [46] includes 32k video clips of 10-second length from 36 musical instrument classes, a subset of VGG-Sound [13], and each video only has one single instrument class annotation. MUSIC [22] consists of 448 untrimmed YouTube music videos of solos and duets from 11 instrument categories, where we use 358 solo videos for training and 90 solo videos for evaluation.

Table 2: **Audio-visual classification** on VGGSound-Music, VGGSound-All, and AudioSet datasets.

Method	VGGSound-Music		VGGSound-All		AudioSet	
	Linear (%)	Finetune (%)	Linear (%)	Finetune (%)	Linear (%)	Finetune (%)
MAE [51]	25.32	52.39	15.61	45.73	11.52	24.23
AudioMAE [47]	41.65	55.61	42.35	57.76	30.23	44.92
CAV-MAE [48]	60.53	67.26	55.27	65.53	40.56	51.29
MAViL [49]	61.95	69.53	57.36	67.17	43.62	53.38
AV-MAE [50]	60.82	67.61	56.15	65.08	41.67	51.32
ACAV-1M (ours)	**64.87**	**71.25**	**61.35**	**69.29**	**47.83**	**56.05**

Here, we explain the baselines, task-specific methods, pre-trained models, and multimodal foundation used in our audio-visual benchmark.

Task-specific methods. *ACAV-1M* is utilized to establish a variety of task-specific methods tailored to each downstream task. For Audio-Visual Classification, methods are optimized for maximum accuracy. In Audio-Visual Source Localization, algorithms focus on improving the mean Intersection over Union (mIoU). For Audio-Visual Retrieval, the emphasis is on enhancing recall rates. Similarly, task-specific approaches are devised for Audio-Visual Video Parsing, Scene-Aware Dialog, Question-Answering, Segmentation, and Source Separation, each aiming to excel in metrics such as F-score, BLEU, METEOR, and Signal Decomposition Ratings (SDR, SIR, SAR).

Pre-trained models. We evaluate several models pre-trained on *ACAV-1M* including audio-MAE [47], CAV-MAE [48], MAViL [49], and AVMAE [50]. These models leverage masked autoencoding techniques tailored for either audio alone or audio-visual data on audio-visual classification to have a comprehensive understanding on these models.

Our method. We use audio-visual masked autoencoders [51, 47] with masked modeling objectives. Specifically, we apply a modality-specific encoder with self-attention transformers to encode unmasked patches and use a decoder to predict the masked patches of the input modality from unmasked encoded and masked tokens. The overall model is simply optimized to reconstruct the original input modality of masked tokens using a ℓ-2 norm objective across predicted audio/visual tokens $\hat{\mathbf{x}}_m^a, \hat{\mathbf{x}}_m^v$ and ground-truth tokens $\mathbf{x}_m^a, \mathbf{x}_m^v$ defined as:

$$\mathcal{L} = \frac{1}{M^a}\sum_{m=1}^{M^a} ||\mathbf{x}_m^a - \hat{\mathbf{x}}_m^a||_2^2 + \frac{1}{M^v}\sum_{m=1}^{M^v} ||\mathbf{x}_m^v - \hat{\mathbf{x}}_m^v||_2^2, \tag{1}$$

where M^a, M^v denote sets of random masks applied on the input patch embeddings for audio and visual tokens, separately.

4 Experiments

4.1 Experimental Setup

Datasets. We use audio-visual pairs from our *ACAV-1M* dataset for pre-training. We finetune the model on datasets specific to the downstream tasks, as described in Section 3.2.

Evaluation metrics. Following the prior work [46, 52, 42], we use the Precision and F1 scores defined in [42] for visual source localization. For source separation, following [22], we use Signal-to-Distortion Ratio (SDR) and Signal-to-Artifact Ratio (SAR). For audio-visual segmentation, we apply mIoU and F1 scores as evaluation metrics, following the previous work [19]. Linear probing and fine-tuning classification evaluations are based on top-1 accuracy, which measures the class difference from the ground-truth labels. For video parsing, we use F-scores to evaluate segment-level predictions for audio-visual events and Type@AV & Event@AV for the overall evaluation performance.

Implementation. The input images are resized to 224 × 224. The audio is represented by log spectrograms extracted from $10s$ of audio at a sample rate of 8000Hz. We follow the prior work [52] and apply STFT to generate an input tensor of size 128 × 128 (128 frequency bands over 128 timesteps) using 50ms windows with a hop size of 25ms. For the audio and visual encoder, we use single-modality MAEs [51, 47]. The models were trained on four A100 GPUs for 100 epochs using the Adam optimizer [53] with a learning rate of $1e-4$ and a batch size of 128.

Table 3: **Audio-visual source localization.** Quantitative results on Flickr-SoundNet.

Method	Precision	AP	F1
Attention 10k [7]	49.38	51.23	55.39
OTS [54]	51.23	53.28	58.12
DMC [30]	50.52	52.93	57.56
CoarsetoFine [55]	51.76	54.85	58.63
DSOL [56]	55.29	57.92	62.05
LVS [57]	52.38	55.31	59.35
EZVSL [52]	54.71	57.51	61.38
Mix-and-Localize [46]	55.83	58.21	62.52
SLAVC [42]	55.65	58.12	62.39
ACAV-1M (ours)	**58.67**	**60.75**	**65.02**

Table 4: **Audio-video retrieval.** Quantitative results on MSR-VTT dataset.

Method	R@1	R@5	R@10
AVLnet [58]	19.62	50.32	60.51
TVLT [59]	23.83	52.56	63.92
ACAV-1M (ours)	**26.57**	**58.78**	**70.26**

Table 5: **Audio-visual scene-aware dialog.** Quantitative results on DSTC10 dataset.

Method	BLEU	METEOR
MMA [60]	24.91	19.36
BMT [61]	36.23	22.83
JSTL [44]	38.52	24.71
ACAV-1M (ours)	**43.27**	**28.65**

4.2 Benchmark Experimental Results

Audio-Visual Classification. To validate the effectiveness of *ACAV-1M* on audio-visual classification, we compare to the following prior baselines: 1) MAE [51]: a masked autoencoder with only images as input; 2) AudioMAE [47]: a masked autoencoder with only audio as input; 2) Audio-Visual MAEs [48, 49, 50]: masked autoencoders with both audio and images as input. Table 2 reports the quantitative comparison results. On VGGSound-Music, we achieved top results with 64.87% in linear probing and 71.25% in fine-tuning, indicating robustness in music-specific scenes. For VGGSound-All, we also recorded 61.35% in linear probing and 69.29% in fine-tuning, showcasing versatility across diverse audio-visual contexts. Regarding AudioSet, our model performed well with 47.83% in linear probing and 56.05% in fine-tuning, reflecting strong generalization capabilities.

Audio-Visual Source Localization. To validate the effectiveness of the proposed *ACAV-1M* dataset for sound source localization, we compare to the following prior work: 1) Attention 10k [7] (CVPR 2018): the first baseline on sound source localization using a two-stream and attention-based neural network; 2) OTS [54] (ECCV 2018): a correspondence-based baseline for localization; 3) DMC [30] (CVPR 2019): a deep multi-modal clustering approach based on audio-visual co-occurrences; 4) CoarsetoFine [55] (ECCV 2020): a two-stage approach using coarse-to-fine embedding alignment; 5) DSOL [56] (NeurIPS 2020): a class-based method with two-stage training; 6) LVS [57] (CVPR 2021): a contrastive learning framework with hard negative mining to learn audio-visual correspondence maps; 7) EZ-VSL [52] (ECCV 2022): a recent weakly supervised localization framework based on multiple-instance contrastive learning; 8) Mix-and-Localize [46] (CVPR 2022): a recent method based on a contrastive random walk on a graph of images and separated sound sources. 9) SLAVC [42] (NeurIPS 2022): a strong baseline with momentum encoders and extreme visual dropout to identify negatives and solve significant overfitting. The results are reported in Table 3. As can be seen, our *ACAV-1M* scored 58.67% Precision, which is the highest among the compared methods, indicating a high accuracy in predicting the correct localization of sound sources. We also achieved 60.75% Average Precision (AP), highlighting the method's consistent performance across different thresholds, outperforming other methods in handling diverse scenarios. Our model achieves a 65.02% F1 score, which reflects the balance between precision and recall, demonstrating the robustness of our approach to effectively localize sound sources.

Audio-Visual Retrieval. For audio-visual retrieval, we evaluated the performance of our *ACAV-1M* model against established methodologies. This evaluation was performed using the MSR-VTT dataset, a comprehensive and challenging benchmark for video understanding and retrieval tasks, where we compare to the following baselines: 1) AVLnet [58]: A self-supervised learning approach that develops a joint audio-visual-textual embedding space, leveraging the natural synchrony in videos to align raw video, audio, and text signals without requiring manual annotations. 2) TVLT [59]: A very recent approach that introduces a visual-audio pre-training framework and incorporates masked audio/video autoencoding coupled with contrastive modeling, which aims to fine-tune the alignment between video and audio modalities to improve retrieval accuracy. The experimental results are shown in Table 4. In particular, we achieved 26.57% R@1, significantly higher than AVLnet (19.62% R@1) and TVLT (23.83% R@1), indicating a more precise retrieval at the topmost rank. Meanwhile,

Table 6: **Audio-visual video parsing.** Quantitative results on LLP dataset.

Method	Audio-Visual	Type@AV	Event@AV
AVE [21]	35.43	39.92	41.63
AVSDN [31]	37.12	45.73	50.82
HAN [16]	48.92	54.03	55.42
MGN [41]	50.63	55.62	57.25
ACAV-1M (ours)	**55.35**	**58.96**	**58.67**

Table 7: **Audio-visual question answering.** Quantitative results on MUSIC-AVQA.

Method	Audio	Visual	Audio-Visual
AVSD [62]	68.52	70.83	65.49
Pano-AVQA [63]	70.73	72.56	66.64
AVQA [18]	74.06	74.00	69.54
ACAV-1M (ours)	**76.87**	**76.65**	**73.25**

Table 8: **Audio-visual segmentation.** Quantitative results on AVSBench dataset.

Method	mIoU	F1
Attention 10k [7]	20.76	31.25
OTS [54]	24.55	36.85
DMC [30]	23.51	35.27
CoarsetoFine [55]	26.53	38.62
DSOL [56]	29.85	42.23
LVS [57]	27.32	40.18
EZVSL [52]	30.52	43.26
Mix-and-Localize [46]	31.69	45.35
SLAVC [42]	31.36	45.02
ACAV-1M (ours)	**36.39**	**49.85**

we scored 58.78% R@5, surpassing both AVLnet (50.32% R@5) and TVLT (52.56% R@5). Our *ACAV-1M* model demonstrates superior performance across all recall metrics.

Audio-Visual Scene-Aware Dialog. In the task of audio-visual scene-aware dialog, model are evaluated to demonstrate their capability to generate contextually appropriate dialog based on both visual and auditory inputs. We compared the performance against several prominent methods in the field: 1) MMA [60]: an end-to-end conversation model that generates dialog responses based on multimodal attention-based video features, which integrates audio and visual cues to form a comprehensive understanding of the video content. 2) BMT [61]: a bi-modal Transformer that adapts the traditional Transformer architecture for bi-modal inputs, processing both audio and visual modalities to enhance performance on tasks like dense video captioning. 3) JSTL [44]: a recent AV-transformer that employs attentional multimodal fusion and combines joint student-teacher learning and model combination techniques to refine dialog generation based on audio-visual data. Table 5 reports the results on the DSTC10 dataset. We observe a BLEU score of 43.27, surpassing all other compared models and reflecting its superior ability to generate grammatically and semantically correct sentences. With a METEOR score of 28.65, our model also leads in this metric.

Audio-Visual Video Parsing. In audio-visual video parsing, we conducted a comparative analysis using the LLP dataset. The comparison the following approaches: 1) AVE [21]: An audio-guided co-attention network which includes additional branches for audio-visual parsing. This model leverages audio cues to enhance the segmentation and identification of visual elements in video. 2) AVSDN [31]: A dual sequence-to-sequence model that merges global audio-visual features into localized contexts. This model aims to improve the parsing accuracy by enhancing the interaction between audio and visual modalities. 3) HAN [16]: A hybrid attention network that utilizes multimodal multiple instance learning pooling. This network focuses on capturing the intricate relationships between audio and visual cues within video content to refine parsing accuracy. 4) MGN [41]: A Multi-modal Grouping Network that aggregates event-aware unimodal features through semantically-aware grouping. It employs learnable categorical embedding tokens. Table 6 shows the experimental results. We achieved an accuracy of 55.35 and a Type@AV score of 58.96, the highest among all compared models, showcasing its exceptional ability to classify and understand different types of content accurately. Furthermore, we scored an Event@AV score of 58.67, illustrating strong performance in identifying and segmenting specific events within videos.

Audio-Visual Question-Answering. For audio-visual question answering (AVQA), our model was assessed on the MUSIC-AVQA dataset, testing its capability to integrate and interpret audio, visual, and combined audio-visual information to answer related questions accurately. This performance was benchmarked against the following models: 1) AVSD [62]: a straightforward approach for audio-visual scene-aware dialog, trained end-to-end to tackle AVQA by directly associating audio-visual scenes with dialog responses. 2) Pano-AVQA [63]: a multimodal transformer encoding with a unique approach to attention mechanisms that incorporate both audio and visual inputs simultaneously. 3) AVQA [18]: a very recent baseline that integrates comprehensive multimodal information by associating spatial grounding, temporal grounding, and advanced multimodal fusion techniques. Table 7 illustrates the experimental results on the MUSIC-AVQA dataset. For instance, we achieved

Table 9: **Sound source separation.** Quantitative results on MUSIC and VGGSound datasets.

Method	MUSIC		VGGS-Instruments		VGGS-Music	
	SDR	SAR	SDR	SAR	SDR	SAR
NMF [64]	-0.62	2.41	-3.85	-0.76	-7.12	-9.01
RPCA [65]	0.86	3.81	-2.39	1.58	-5.53	-7.82
Sound-of-Pixels [22]	4.55	10.24	2.52	4.67	0.95	1.03
MP-Net [66]	4.82	10.56	2.63	4.85	1.37	1.39
CCoL [67]	6.35	9.75	3.28	5.01	2.07	2.18
OneAVM [45]	7.38	7.48	5.36	5.52	2.51	2.61
ACAV-1M (ours)	**10.75**	**11.23**	**8.23**	**8.38**	**5.06**	**5.32**

Table 10: **Ablation results on the benefit of our data curation pipeline** across different audio-visual benchmarks. Note that we use 100K VGGSound samples for pre-training.

Data Curation	Cls. Acc (%)	SrcLoc. Prec	Retrieval Acc (%)	SADialog. BLEU	VidPars. F-score (%)	QA Acc (%)	Seg. mIoU	SrcSep. SDR
✗	36.82	45.29	8.79	31.57	38.73	55.32	21.38	3.52
✓	**45.38**	**49.72**	**15.56**	**34.83**	**41.96**	**60.82**	**24.62**	**4.63**

a 76.87%@Audio score, indicating a high proficiency in extracting and utilizing audio information to answer questions, outperforming all other models. With a score of 73.25%@Audio-Visual, our model demonstrates a superior ability to use audio and visual data for answering questions, surpassing other methodologies in effectively utilizing integrated multimodal cues.

Audio-Visual Segmentation. In audio-visual segmentation, we did a comparative analysis using the AVSBench [19] dataset, which is designed to evaluate segmentation capabilities across models that integrate audio and visual data. This task extends beyond localization to include the generation of accurate segmentation masks for audio-visual sources. We use the same baselines [7, 54, 30, 55, 56, 57, 52, 46, 42] as those for audio-visual source localization, adapted to generate detailed segmentation masks rather than just coarse localization maps. The results are reported in Table 8. Our model achieved a mIoU score of 36.39, indicating superior accuracy in segmenting relevant audio-visual content precisely. We also achieved an F1 score of 49.85, the highest among all compared methods. These results highlight the *ACAV-1M* model's robust capability to accurately segment complex audio-visual scenes, establishing it as a leading method for audio-visual segmentation.

Audio-Visual Source Separation. To demonstrate the effectiveness of the proposed *ACAV-1M* on source separation, we compare to the following methods: 1) NMF [64]: a traditional signal processing approach based on non-negative matrix factorization to generate the spectrogram of each sound source; 2) RPCA [65]: a parameter-free baseline based on robust principal component analysis; 3) Sound-of-Pixels [22]: a deep learning approach that recovers separated audio conditioned on pixel-level visual features; 4) MP-Net [66]: an improved audio-visual method based on recursive separation from the mixture; 5) CCoL [67] (CVPR 2021): a cyclic co-learning framework based on sounding object visual grounding to separate individual sound sources. 6) OneAVM [45] (ICML 2023): a unified audio-visual framework for localization, separation, and recognition. We report the comparison results in Table 9. Our model showcased the best performance in both SDR and SAR across all datasets. For example, we achieved an SDR score of 10.75 on the MUSIC dataset, significantly higher than other methods, reflecting superior separation quality. Meanwhile, our model also reached an SDR score of 8.23 and 5.06 on VGGSound-Instruments and VGGSound-Music, respectively. These results underscore the effectiveness of *ACAV-1M* for audio-visual source separation.

4.3 Experimental Analysis

In this section, we provide a detailed analysis of our experimental results to demonstrate the benefits of our data curation pipeline, the impact of the quality measure criteria on model performance, and the scaling behavior of the *ACAV-1M* dataset.

Benefit of data curation pipeline. To demonstrate the efficacy of our data curation pipeline, we conducted comparative experiments using a random subset of VGGSound with equivalent size and a clean subset of VGGSound with our data curation pipeline. The experimental results are reported in Table 10. The results clearly indicate that improvements are attributed to our data curation process, which includes precise alignment of audio and visual data and careful annotation. The clean subset

Table 11: **Ablation results on alignment filtering** across different audio-visual benchmarks.

Quality Measure	Cls. Acc (%)	SrcLoc. Prec	Retrieval Acc (%)	SADialog. BLEU	VidPars. F-score (%)	QA Acc (%)	Seg. mIoU	SrcSep. SDR
–	33.25	40.67	4.86	29.78	36.93	50.19	15.86	1.87
Instance Align	39.56	43.95	6.17	31.23	38.21	53.27	18.37	2.38
+ Temporal Align	42.68	46.53	12.65	33.15	40.16	55.32	20.65	3.29
+ Language Align	**47.85**	**50.27**	**16.78**	**35.34**	**42.65**	**60.26**	**24.83**	**4.35**
threshold=70%	45.76	49.12	15.69	34.87	42.36	59.75	24.02	4.13
threshold=50%	44.65	48.35	14.73	34.06	41.89	58.23	22.96	3.67

Table 12: **Ablation results on the scaling property of our ACAV-1M** across different audio-visual benchmarks.

Data Scale	Cls. Acc (%)	SrcLoc. Prec	Retrieval Acc (%)	SADialog. BLEU	VidPars. F-score (%)	QA Acc (%)	Seg. mIoU	SrcSep. SDR
10K	30.63	37.82	4.23	28.56	33.25	45.63	11.56	1.25
52K	47.85	50.27	16.78	35.34	42.65	60.26	24.83	4.35
100K	49.27	51.85	17.63	36.08	43.72	62.23	25.98	5.12
199K	51.73	53.62	19.35	38.15	46.58	65.73	28.75	6.28
1M	**61.35**	**60.75**	**26.57**	**43.27**	**55.35**	**73.25**	**36.39**	**10.75**

of VGGSound with our data curation pipeline achieves superior performance and demonstrates the value of high-quality, well-aligned data in training more effective multimodal models.

Ablation on alignment filtering. We analyzed the impact of different quality measure criteria on model performance, as shown in Table 11. For language alignment, we experimented with using class names directly instead of our detailed, long captions for annotations. This change resulted in a noticeable degradation in performance, emphasizing the importance of rich, descriptive captions in providing contextual cues that enhance model understanding and performance. Regarding temporal alignment, we varied the alignment accuracy of audio and visual data during the dataset curation process, testing alignment accuracies of 100%, 70%, and 50%. Our experiments show that models trained with 100% alignment accuracy consistently outperform those trained with lower accuracies, underscoring the critical role of precise synchronization in audio-visual learning.

Scaling property of our ACAV-1M dataset. To understand the scalability of our dataset, we trained models on progressively larger subsets of *ACAV-1M*, specifically 10K, 52K, 100K, 199K, and 1M samples. The experimental results across different downstream tasks are reported in Table 12. Our findings reveal a positive correlation between the size of the dataset and the performance. As the size increases, we observe improved accuracy and robustness across all tasks, indicating that *ACAV-1M* not only supports effective training at smaller scales but also benefits significantly from scaling up.

5 Conclusion

In this work, we introduce *ACAV-1M*, a novel large-scale dataset with one million samples that are curated to bridge the gap between audio and visual data. Furthermore we propose a comprehensive audio-visual benchmark that supports a wide array of audio-visual tasks, from classification and segmentation to retrieval and scene-aware dialog, each benefiting from the dataset's rich annotations and precise audio-visual alignment. We demonstrate the superior performance of models trained on *ACAV-1M* compared to existing methods. Our experiments also explore the scaling behavior of the dataset, showing significant improvements in model performance as data volume increases, thus confirming the dataset's scalability.

Limitations and broader impact. While our dataset covers a broad range of audio and visual contexts, there are still rare scenarios that are underrepresented or absent, which could affect the generalizability of the trained models to all real-world applications.

ACAV-1M has the potential to make a profound impact for audio-visual learning. The insights gained from models trained on *ACAV-1M* can enhance multimedia applications, improve accessibility features, and foster the development of intuitive and interactive systems. However, it is essential to be aware of ethical considerations and potential biases in training data, which could amplify disparities if not carefully managed.

References

[1] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023. 1, 3

[2] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023. 1, 3

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 1, 2, 3

[4] Andrew Owens, Jiajun Wu, Josh H. McDermott, William T. Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–816, 2016. 1, 3

[5] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 1, 3

[6] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 1, 3

[7] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018. 1, 3, 4, 6, 7, 8

[8] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 1, 3

[9] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 4

[11] 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 International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[12] Jort F. Gemmeke, Daniel P. W. 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 *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 2, 3

[13] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 2, 3, 4

[14] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019. 2

[15] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 2

[16] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 436–454, 2020. 2, 3, 4, 7

[17] Huda AlAmri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Peter Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. Audio visual scene-aware dialog. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7550–7559, 2019. 2

[18] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19086–19096, 2022. 2, 4, 7

[19] 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 *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 2, 4, 5, 8

[20] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740, 2019. 2

[21] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018. 2, 3, 7

[22] 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. 2, 3, 4, 5, 8

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

[24] 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. 3

[25] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 4733–4744, 2020. 3

[26] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12934–12945, 2021. 3

[27] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12475–12486, June 2021. 3

[28] John Hershey and Michael Casey. Audio-visual sound separation via hidden markov models. *Advances in Neural Information Processing Systems*, 14, 2001. 3

[29] 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. 3

[30] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9248–9257, 2019. 3, 6, 7, 8

[31] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2002–2006, 2019. 3, 7

[32] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6291–6299, 2019. 3

[33] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 3

[34] Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[35] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 324–333, 2019. 3

[36] Changan Chen, Unnat Jain, Carl Schissler, S. V. A. Garí, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 17–36, 2020. 3

[37] Changan Chen, Sagnik Majumder, Al-Halah Ziad, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021. 3

[38] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022. 3

[39] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1326–1335, 2021. 3

[40] Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[41] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 7

[42] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 4, 5, 6, 7, 8

[43] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016. 4

[44] SAnkit Shah, Shijie Geng, Peng Gao, Anoop Cherian, Takaaki Hori, Tim K. Marks, Jonathan Le Roux, and Chiori Hori. Audio-visual scene-aware dialog and reasoning using audio-visual transformers with joint student-teacher learning. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7732–7736. IEEE, 2022. 4, 6, 7

[45] Shentong Mo and Pedro Morgado. A unified audio-visual learning framework for localization, separation, and recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. 4, 8

[46] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10483–10492, 2022. 4, 5, 6, 7, 8

[47] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Proceedings of Advances In Neural Information Processing Systems (NeurIPS)*, 2022. 5, 6

[48] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *Proceedings of The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 5, 6

[49] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Haoqi Fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, and Christoph Feichtenhofer. Mavil: Masked audio-video learners. *arXiv preprint arXiv:2212.08071*, 2022. 5, 6

[50] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16144–16154, October 2023. 5, 6

[51] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 5, 6

[52] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 218–234, 2022. 5, 6, 7, 8

[53] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[54] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018. 6, 7, 8

[55] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 292–308, 2020. 6, 7, 8

[56] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 10077–10087, 2020. 6, 7, 8

[57] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16867–16876, 2021. 6, 7, 8

[58] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020. 6

[59] Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal. Tvlt: Textless vision-language transformer. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 6

[60] Chiori Hori, Huda Alamri, Jue Wang, Gordon Winchern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. *arXiv preprint arXiv:1806.08409*, 2018. 6, 7

[61] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *British Machine Vision Conference (BMVC)*, 2020. 6, 7

[62] Idan Schwartz, Alexander G. Schwing, and Tamir Hazan. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[63] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360° videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 7

[64] 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. 8

[65] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 57–60, 2012. 8

[66] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 8

[67] 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 (CVPR)*, pages 2745–2754, 2021. 8

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We have read the guidelines.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 4.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 4.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.
   (b) Did you mention the license of the assets? [Yes] See Section 4.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include it in the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 4.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 4.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]