ACAV-1M: Data Curation and Benchmarking for Audio-Visual Representation Learning (Supplementary Material)

Anonymous Author(s)
Affiliation
Address
email

In this supplementary material, we provide the following material:

- dataset documentation and intended uses in Section 1,
- dataset website in Section 2,
- croissant metadata in Section 3,
- author statement in Section 4,
- hosting, licensing, and maintenance plan in Section 5,
- algorithm for our data curation pipeline in Section 6,
- addition implementation and datasets details in Section 7.

1 Dataset Documentation & Intended Uses

The *ACAV-1M* dataset is designed to facilitate research in audio-visual representation learning. It includes synchronized audio and visual data curated from various sources to ensure a diverse and comprehensive collection for training and evaluating machine learning models. The dataset documentation follows the *datasheets for datasets* framework, providing detailed information on the dataset's composition, collection process, and intended uses.

Composition: The dataset consists of 1 million audio-visual pairs, including video clips with corresponding audio tracks from various domains such as user-generated content.

Collection Process: Data was collected using automated scripts and manual curation to ensure quality and relevance. Metadata includes source URLs, timestamps, and content descriptions.

Intended Uses: The dataset is intended for developing and benchmarking models in audio-visual representation learning, including tasks like video classification, audio-visual synchronization, and cross-modal retrieval.

Ethical Considerations: We ensured that the dataset adheres to ethical guidelines, including the exclusion of sensitive or inappropriate content and respect for copyright and privacy concerns.

This document is based on *Datasheets for Datasets* by Gebru *et al.* [1].

MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

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

The *ACAV-1M* was created to fill a significant gap in multimodal learning where audio and visual data are integrated systematically. It aims to enhance robust models that leverage both modalities for improved understanding and interaction, designed specifically for tasks like audio-visual classification, localization, retrieval, and segmentation.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
The dataset was created by a collaborative effort involving researchers from various academic institutions specializing in machine learning and computer vision, under the coordination of a leading university's computer science department.

What support was needed to make this dataset? (e.g.who funded the creation of the dataset? If there is an associated grant, provide the name of the grantor and the grant name and number, or if it was supported by a company or government agency, give those details.)
No. The creation of the *ACAV-1M* was not supported by any grants from several research funding agencies. However, the dataset development received technical support and infrastructure from the host university.

Any other comments?
No.

COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
The instances in the *ACAV-1M* represent synchronized audio-visual clips from diverse settings, including music performances, public speeches, and everyday activities, ensuring a wide range of scenarios for robust multimodal learning.

How many instances are there in total (of each type, if appropriate)?
The dataset comprises approximately 100,000 video clips, each paired with corresponding audio tracks that have been meticulously synchronized and annotated.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).
The dataset is a curation subset of the original ACAV [2] dataset with 100 million samples.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Each instance consists of "Raw" video and audio data. Additional metadata include synchronization points, annotations for source localization, and labels for classification and segmentation tasks.

Is there a label or target associated with each instance? If so, please provide a description.
Yes, each instance includes captions associated with the video and audio. For various tasks, we include labels for each instance like classification (audio-visual context), segmentation masks, and

localization coordinates.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.
No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.
No.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.
No.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
No.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.
Yes. The dataset is a curation subset of the original ACAV [2] dataset with 100 million samples.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.
No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.
No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.
No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.
No.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No.

Any other comments?

No.

COLLECTION

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was acquired through a combination of public domain resources and contributions from collaborating institutions, where scenarios were staged and recorded under controlled conditions to ensure quality and diversity.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. Finally, list when the dataset was first published.

Data collection spanned over half one year, culminating in the dataset's release in 2024. The temporal alignment of collection and creation ensured the relevance and recency of the data.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

We have alignment filtering mechanisms to curate our dataset from the original ACAV [2] dataset.

What was the resource cost of collecting the data? (e.g. what were the required computational resources, and the associated financial costs, and energy consumption - estimate the carbon footprint. See Strubell *et al.*[3] for approaches in this area.)

We use four A100 GPUs to curate data and train our models.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

We used alignment filtering mechanisms.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Authors are involved in the data curation process.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
No.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
No.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
No.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.
No.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate)
No.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
No.

Any other comments?
No.

PREPROCESSING / CLEANING / LABELING

Was any preprocessing/cleaning/labeling of the data done(e.g.,discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
Yes. We use multimodal LLM to .

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
No.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

No.

Any other comments?
No.

USES

Has the dataset been used for any tasks already? If so, please provide a description.
Yes, *ACAV-1M* has been employed in several benchmarking tasks within the research group, including preliminary studies on audio-visual perception tasks.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.
No.

What (other) tasks could the dataset be used for?
Beyond the current uses, the dataset holds potential for tasks in automated content generation, assistive technologies, and advanced surveillance systems.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?
No.

Are there tasks for which the dataset should not be used? If so, please provide a description.
No.

Any other comments?
No.

DISTRIBUTION

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
No.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
The dataset is available via a website page and can be accessed through the dataset page, which ensures controlled and ethical usage aligned with academic standards.

When will the dataset be distributed?
The dataset will be available upon publication.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

No.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

Any other comments?

No.

MAINTENANCE

Who is supporting/hosting/maintaining the dataset?

The dataset is maintained by the authors, with plans for ongoing updates and expansions based on community feedback and technological advancements.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

The owner of the dataset can contacted by email.

Is there an erratum? If so, please provide a link or other access point.

No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes, the dataset is scheduled for regular reviews and updates to address any errors, introduce new instances, and phase out obsolete data, with all changes communicated through the dataset's official repository.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes. It will be maintained on the dataset website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.
Yes. We will open the opportunity for other researchers to augment the dataset for additional benchmarks.

Any other comments?
No.

2 Dataset Website

The dataset and its documentation can be accessed at the following URL:

https://acav1m.github.io

This website provides an overview of the dataset, download links, and additional resources such as example code, tutorials, and a forum for community discussions. Users can explore the dataset through an interactive interface, which includes search and filter options to facilitate easy access to specific subsets of the data.

3 Croissant Metadata

The Croissant metadata for the ACAV-1M dataset is available at:

https://acav1m.github.io

This metadata record documents the dataset's structure, including descriptions of the files, their formats, and the fields within each record. The metadata adheres to the Croissant format, ensuring interoperability and ease of use with ML tools and platforms.

4 Author Statement

We, the authors of the ACAV-1M dataset, bear full responsibility for any violations of rights and confirm that all data included in the dataset complies with the relevant licenses and ethical guidelines. The dataset is released under the Creative Commons Attribution 4.0 International License (CC BY 4.0), which allows for sharing, adaptation, and use of the data with appropriate credit given to the original authors.

5 Hosting, Licensing, & Maintenance Plan

The dataset is hosted on a dedicated server managed by our institution, ensuring reliable access and download speeds. We also provide mirror links through major cloud storage providers to ensure redundancy and availability. The dataset is licensed under the Creative Commons Attribution 4.0 International License.

6 Pseudo Algorithm for ACAV-1M Data Curation and Alignment

Algorithm 1 is a pseudo-algorithm that encapsulates the data curation and alignment filtering processes described for the *ACAV-1M* dataset. This algorithm is structured to provide a clear, step-by-step procedure that reflects the robust methodologies used in preparing the dataset. This algorithm also provides a structured approach to processing and aligning the data within the *ACAV-1M* ensuring that each component (video, audio, and textual caption) is effectively synchronized and semantically

Algorithm 1 Pseudo Algorithm for Data Curation and Alignment Filtering

Input: Raw video and audio data
Output: Curated dataset with aligned audio-visual captions

Data Curation Process:
for each video clip in dataset **do**
    Extract raw audio and video streams
    Use VideoLLaVA [4] to generate sentence-level descriptions from video
    Condense descriptions using a general LLM [5] into a single comprehensive caption
    Attach the comprehensive caption to the corresponding video clip
end for

Alignment Filtering Process:
for each item in curated dataset **do**
    Language Alignment:
    Calculate normalized cosine similarity between text captions and audio-visual content
    if similarity < 0.5 **then**
        Flag for review or reprocessing
    end if
    Instance Alignment:
    Assess synchronization between audio and visual streams using ImageBind [6]
    if similarity < 0.5 **then**
        Adjust synchronization parameters and re-align
    end if
    Temporal Alignment:
    Check for alignment within a temporal window of 1 second per segment
    if average alignment threshold < 0.5 **then**
        Refine temporal synchronization parameters
    end if
end for

Return finalized dataset with validated and aligned captions

coherent. The algorithm is designed to be part of a larger document or paper, offering clarity on the methods and steps taken to curate and align data within the dataset.

7 Implementation & Dataset Details

In this section, we provide more implementation and dataset details.

Audio-visual classification. For linear probing, we follow the prior work [7, 8] and extract frozen audio-visual representations from our *ACAV-1M* pre-trained audio-visual masked autoencoder. Then we attach a linear layer as a head to the frozen features for training with the audio-visual classes. During training, we only fine-tune the linear head to evaluate the quality of pre-trained features. The models are trained for 50 epochs using the Adam optimizer [9] with a learning rate of $1e-4$ and a batch size of 128. For fine-tuning, we use the same optimizer and batch size settings, but all parameters are learnable.

Audio-Visual Source Localization. For sound source localization, we train all baselines [10, 11, 12] using the same backbone (*i.e.*, ViT-Base) for audio/visual encoder with different proposed objectives in their original papers. The final localization map is generated through bilinear interpolation of the similarity map between audio/visual features from the last self-attention layer. The models are trained for 30 epochs using the Adam optimizer [9] with a learning rate of $1e-4$ and a batch size of 128.

Audio-Visual Retrieval. The retrieval task processes video frames sampled at 8 fps and utilizes combined low-level visual features from ResNet-152 [13] and 3D ResNet models [14], both pre-trained on respective large-scale datasets. Audio features are extracted using VGGish [15], pre-trained

on AudioSet [16]. The complete model, integrating these features, is trained using Adam to optimize retrieval effectiveness across 40 epochs.

Audio-Visual Video Parsing. Following the data pre-processing in previous work [17], we sample video frames at 8 fps from the 10-second videos with 10 non-overlapping snippets of 1 second. For low-level visual features, we concatenate 2D and 3D visual features extracted by ResNet-152 [13] pre-trained on ImageNet [18] and 3D ResNet [14] pre-trained on Kinetics-400 [19]. We utilize VGGish [15] pre-trained on AudioSet [16] to extract the audio features. The model is trained with Adam [9] optimizer with β_1=0.9, β_2=0.999 and with an initial learning rate of 3e-4. We train the model with a batch size of 16 for 40 epochs. Note that each video includes at least 1s audio or visual event, and 7202 video clips are annotated with more than one event category. We use 10,000 video clips with only video-level event labels for training. Following the official splits [17] of validation and test sets, we develop and test the model on the remaining 1879 videos with the segment-level annotations, *i.e.*, the speech event for audio starts at 1s and ends at 5s.

Audio-Visual Scene-Aware Dialog. In the audio-visual scene-aware dialog task, our model employs an advanced dialog generation framework that integrates audio and visual information to produce contextually relevant conversations. The dialog system utilizes a Transformer-based architecture, which processes inputs from both modalities through separate encoders before merging them in a fusion layer. This approach allows the model to understand the context provided by both the audio and visual data streams effectively. The model is optimized using the Adam optimizer with a learning rate of $1e-4$ and a batch size of 64. Training is conducted for up to 30 epochs, with early stopping based on performance on a validation set to prevent overfitting.

Audio-Visual Question-Answering. For the AVQA task, our implementation focuses on integrating spatial and temporal grounding techniques to accurately answer questions based on the video and audio content. The system employs a dual-stream encoder that separately processes visual and audio inputs. The encoded features are then combined using a co-attention mechanism that aligns audio and visual elements relevant to the question context. This integration allows the model to focus on specific segments of audio and video that are crucial for answering the given question. The model is trained using the Adam [9] optimizer with an initial learning rate of $3e-4$, reduced by a factor of 0.1 upon plateauing of validation loss. The system is trained for 40 epochs with a batch size of 32.

Audio-Visual Segmentation. For segmentation, we follow the prior work [20], and apply an upsampling decoder on features from the last self-attention layer to generate the final segmentation mask. We use the binary cross entropy (BCE) loss between the prediction and ground-truth masks for training. The models are trained for 20 epochs using the Adam optimizer [9] with a learning rate of $1e-4$ and a batch size of 128.

Audio-Visual Source Separation. For sound source separation, we follow the previous method [21, 22] and attach an audio U-Net decoder to our pre-trained audio-visual encoders for separating sounds from the mixture. The decoder depth for self-attention layers is 8, and the decoder receives the representations of the audio mixture and the visual embeddings. We also apply multiple transposed convolutions and an output head to predict a time-frequency separation mask. This separation mask is then used to multiply the input mixture STFT to separate the audio. Similarly to [21], the target masks refer to the time-frequency bins where the source is the most dominant component in the mixture. The sound source separation is achieved by optimizing a binary cross-entropy loss over these binary targets. The model is trained for 20 epochs using the Adam optimizer [9] with a learning rate of $1e-4$ and a batch size of 128.

Dataset Details. We evaluated our method using several prominent audio-visual datasets:

- **Flick-SoundNet [23]:** a dataset consisting of natural soundscapes with associated Flickr images with 4,500 audio-visual pairs for training and testing the model on 250 audio-visual pairs of sounding objects and extended 250 non-sounding objects;

- **VGG-Instruments [12]:** contains video clips of musical instrument performances, with 32k video clips of 10s lengths from 36 musical instrument classes, a subset of VGG-Sound [24], and each video only has one single instrument class;
- **MUSIC [21]**: consists of 448 untrimmed YouTube music videos of solos and duets from 11 instrument categories;
- **VGG-Music [22]:** a dataset that features a collection of music videos with annotations related to the genre and instruments present;
- **VGGSound [24]:** a comprehensive dataset that includes a wide variety of sound categories and corresponding visual scenes, which contains categories, such as animals, instruments, vehicles, people, etc;
- **AudioSet [16]:** a collection of 2,084,320 human-labeled 10-second sound clips drawn from YouTube videos with 632 audio event classes;
- **AVSBench [20]:** a benchmark for testing audio-visual synchronization and alignment in diverse settings, including 4,932 videos (in total 10,852 frames) from 23 categories, including instruments, humans, animals, etc.
- **MSR-VTT [25]:** A large-scale video description dataset that includes 10,000 video clips, each paired with 20 human-annotated captions, useful for tasks involving video understanding and retrieval.
- **LLP [17]:** The Look, Listen, and Parse (LLP) dataset contains densely labeled video segments that are used to train and evaluate models on tasks requiring fine-grained temporal understanding of video content.
- **MUSIC-AVQA [26]:** A dataset specifically curated for audio-visual question answering in 11,849 YouTube video clips of 10 seconds long from 25 different event categories. It combines visual and audio clues to answer complex queries about the content and context of musical pieces.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for Datasets. *arXiv preprint arXiv:1803.09010*, 2018. 1

[2] 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, 4

[3] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and Policy Considerations for Deep Learning in NLP. *arXiv preprint arXiv:1906.02243*, 2019. 4

[4] 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. 9

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

[6] 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. 9

[7] 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. 9

[8] 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. 9

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 9, 10

[10] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 218–234, 2022. 9

[11] 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. 9

[12] 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. 9, 11

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 9, 10

[14] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459, 2018. 9, 10

[15] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. CNN architectures for large-scale audio classification. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 9, 10

[16] 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. 10, 11

[17] 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. 10, 11

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia. Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 10

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

[20] 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. 10, 11

[21] 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. 10, 11

[22] 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. 10, 11

[23] 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. 10

[24] 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. 11

[25] 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. 11

[26] 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. 11