
AVSET-10M: An Open Large-Scale Audio-Visual Dataset with High Correspondence

A Datasheet of AVSET-10M

We present a datasheet [4] for documentation and responsible usage of LeanDojo Benchmark.

A.1 Motivation

- 1. For what purpose was the dataset created?** We have developed the AVSET-10M dataset, a tailored audio-video corresponding dataset, designed to advance audio-visual research by facilitating the exploration of semantic and temporal alignment between audio and video components.
- 2. Who created the dataset and on behalf of which entity?** The AVSET-10M was developed by researchers listed in the author list.
- 3. Who funded the creation of the dataset?** This work is funded by the Zhejiang University.

A.2 Composition

- 1. What do the instance that comprise the dataset represent (e.g., documents, photos, people, countries?)** Each instance consists of a pair of corresponding audio and video samples, along with several associated labels.
- 2. How many instances are there in total (of each type, if appropriate)?** The AVSET-10M dataset contains 10,605,005 samples, of which the AVSET-700K subset includes 727,530 samples.
- 3. Does the dataset contain all possible instances or is it a sample of instances from a larger set?** The dataset contains all possible instances.
- 4. What data does each instance consist of?** We provide comprehensive meta-information for each video clip, including the YoutubeID of the video, timestamps for each clip, audio-visual cosine similarity, a flag indicating whether sound separation is required, and relevant text labels. For AVSET-10M (w/o. AVSET-700K), captions and pseudo-labels are included, while AVSET-700K features manual audio labels.
- 5. Is there a label or target associated with each instance?** We provide the cosine similarity between audio and visual content as well as the audio labels for each sample.
- 6. Is any information missing from individual instances?** For some instances filtered from Panda-70M, although the audio and video correspond, it is not able to identify the specific audio pseudo-labels. Note that this does not affect the audio-visual correspondence in our dataset.
- 7. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** N/A
- 8. Are there recommended data splits (e.g., training, development/validation, testing)?** In the AVSET-10M dataset, there are a large number of audio labels, allowing researchers to perform appropriate splits based on these labels. We do not have a recommended data splits.
- 9. Are there any errors, sources of noise, or redundancies in the dataset?** N/A

10. **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** We only provide the download links for the videos, the raw videos need to be downloaded from the YouTube platform.
11. **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)?** The AudioSet [5] and Panda-70M [2] used as the source contains facial videos that may pose a risk of infringement, we will delete the corresponding samples if necessary to avoid potential legal issues.
12. **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** Our data all come from the YouTube platform, which has a detailed data review process to ensure that it does not contain videos that are offensive, insulting, threatening, or might otherwise cause anxiety.
13. **Does the dataset identify any subpopulations (e.g., by age, gender)?** N/A
14. **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?** Individual identities may be identifiable through the video uploader.
15. **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race 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)?** N/A

A.3 Collection Process

1. **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)?** The audio-video similarity is calculated using Imagebind [6], and the audio tags are obtained using PANNs [8].
2. **What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?** All raw video data is sourced from established open-source datasets, and we employ an advanced filtering process to refine these data. The integrity and efficacy of the filtering process for the entire dataset have been thoroughly verified in Section 3.3.
3. **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** Based on the audio-video similarity.
4. **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)?** N/A.
5. **Were any ethical review processes conducted (e.g., by an institutional review board)?** Our data all come from the YouTube platform, which has a detailed data review process to ensure that it does not contain videos that are offensive, insulting, threatening, or might otherwise cause anxiety.

A.4 Preprocessing/cleaning/labeling

1. **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)?** We employ Imagebind [6] to determine the similarity between audio and video, PANNs [8] to classify audio into different categories, and a sound separation model [10] to extract non-speech tracks from the audio.

2. **Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** We provide URLs for all raw videos, allowing researchers to download the videos directly from the YouTube platform.
3. **Is the software that was used to preprocess/clean/label the data available?** ImageBind (<https://github.com/facebookresearch/ImageBind>). PANNS (https://github.com/qiuqiangkong/audioset_tagging_cnn). Sound Separation model (<https://github.com/ZFTurbo/MVSEP-CDX23-Cinematic-Sound-Demixing>).

A.5 Uses

1. **Has the dataset been used for any tasks already?** Yes, we have benchmarked the tasks of visual guided sound separation and audio-video retrieval using the AVSET-10M dataset.
2. **Is there a repository that links to any or all papers or systems that use the dataset?** Yes. Please visit the web page of AVSET-10M (<https://avset-10m.github.io>).
3. **What (other) tasks could the dataset be used for?** Our dataset is designed to facilitate research in video-to-audio generation [9], text-to-audio generation [7], and various other audio-video generation tasks. Additionally, it supports studies in audio-video classification [1], audio-video captioning [12], and other related audio-video understanding tasks.
4. **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 dataset consumer 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 risks or harms (e.g., legal risks, financial harms)?** To enlarge the sample size of non-speech categories, we utilize a sound separation model to process the data. This method may introduce a certain degree of audio distortion. Users can create a distortion-free sample set by using the identifiers provided in the dataset.
5. **Are there tasks for which the dataset should not be used?** N/A.

A.6 Distribution

1. **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?** Yes, the dataset is open to the public.
2. **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** The dataset will be distributed through platforms such as github and hugging face, and the code will be placed on github (<https://avset-10m.github.io/>) and hugging face (<https://huggingface.co/datasets/avset10m/avset10m>).
3. **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** No.
4. **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** No.

A.7 Maintenance

1. **Who will be supporting/hosting/maintaining the dataset?** The first author of this paper.
2. **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** The owner/curator/manager(s) of the dataset can be contacted through chengxize@zju.edu.cn
3. **Is there an erratum?** No. If errors are found in the future, we will release errata on the main web page for the dataset <https://avset-10m.github.io/>.
4. **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** Yes, the datasets will be updated whenever necessary to ensure accuracy, and

132 announcements will be made accordingly. These updates will be posted on the main web
133 page for the dataset <https://avset-10m.github.io/>.

134 5. **If the dataset relates to people, are there applicable limits on the retention of the data**
135 **associated with the instances (e.g., were the individuals in question told that their data**
136 **would be retained for a fixed period of time and then deleted?)** The samples in the
137 dataset are sourced from the YouTube platform. We have stated that if any specific fragments
138 are found to infringe on individual rights, we will promptly remove them.

139 6. **Will older version of the dataset continue to be supported/hosted/maintained?** Yes,
140 older versions of the dataset will continue to be maintained and hosted.

141 7. **If others want to extend/augment/build on/contribute to the dataset, is there a mecha-**
142 **nisms for them to do so?** Our dataset will be published on the GitHub platform. If other
143 researchers wish to further expand the dataset, they are welcome to contact us.

144 B Implementation Details

145 B.1 Sound Separation

146 Same as the experimental setting of [3], for all audio samples, we conduct experiments on samples of
147 length 65535 (approximately 4 seconds) at a sampling rate of 16 kHz. For spectrum computation, we
148 employ a short-time Fourier transform (STFT) with a filter length of 1024, a hop length of 256, and
149 a window size of 1024. All images are resized to 224×224 pixels. All models are trained with a
150 batch size of 128, using the Adam optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$,
151 for 200,000 steps. Additionally, we employ warm-up and gradient clipping strategies, following [3].
152 We compute the signal-to-distortion ratio (SDR) using museval [11]. All experiments are conducted
153 on a single A800 GPU.

154 B.2 Audio-Video Retrieval

155 Following the experimental setting of [13], we added a projection layer after each feature extractor to
156 map all representations to the same space. For all experiments, only the projection layer is trainable.
157 The softmax temperature is set to 0.01, and the temperature for the InfoNCE loss is set to 0.02. We
158 utilize the Adam optimizer with a learning rate of 1×10^{-3} and a batch size of 2048, running the
159 training process for 20 epochs on a single A800 GPU.

160 C AVSET-10M

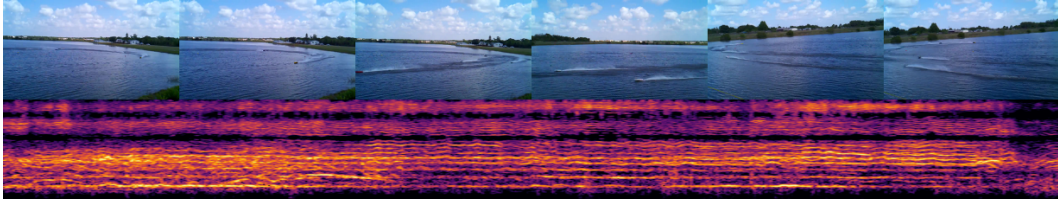
161 C.1 Samples of AVSET-10M

162 We present some audio-video consistency samples from the AVSET-10M in Figure 1. For additional
163 samples, please visit the demo page at <https://avset-10m.github.io> and hugging face page at
164 <https://huggingface.co/datasets/avset10m/avset10m>.

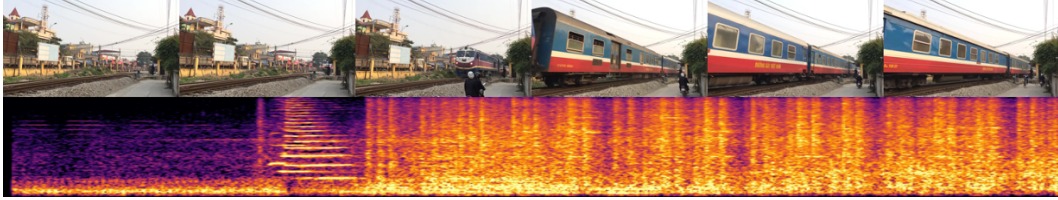
165 C.2 Dataset Composition

166 We release AVSET-10M as the following two subsets:

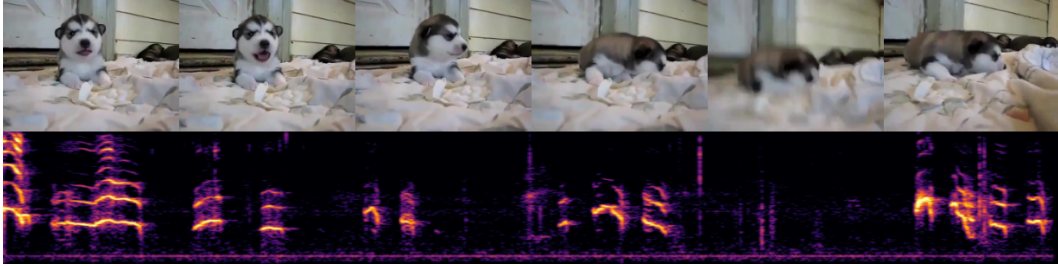
- 167 • **AVSET-700K**: This subset comprises 727,530 audio-visual corresponding samples filtered from
168 AudioSet. Each video segment in this subset is accompanied by a manually labeled audio category,
169 ensuring accurate categorization and relevance.
- 170 • **AVSET-10M (w/o. AVSET-700K)**: This subset comprises 9,877,475 audio-visual corresponding
171 samples, filtered from the Panda-70M dataset. Each video segment is semantically coherent,
172 focusing on a single event, and includes a text description originally from the Panda70M dataset.
173 Additionally, we provide pseudo-labels for the audio categories, derived with PANNs [8], along



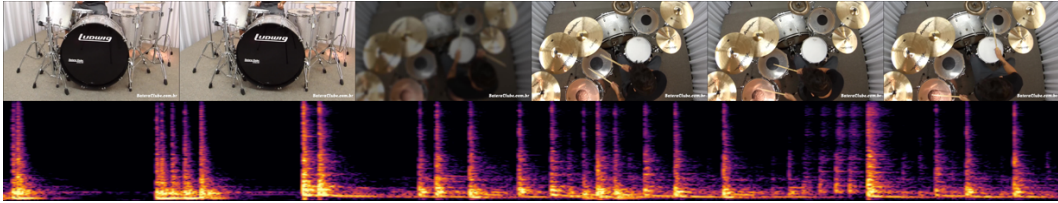
(a) Audio-Vision Cosine Similarity $\theta = 0.479$.



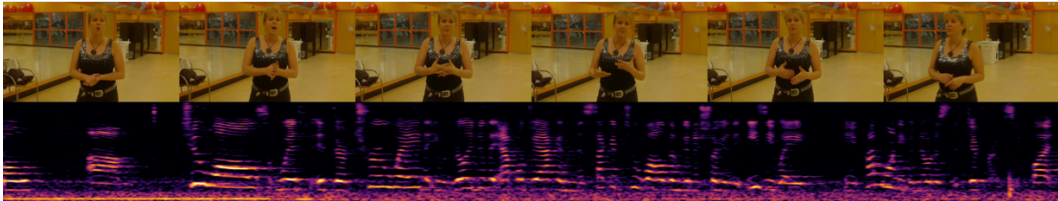
(b) Audio-Vision Cosine Similarity $\theta = 0.442$.



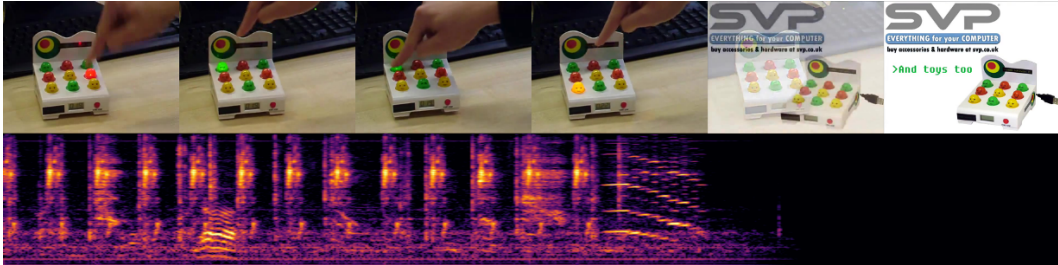
(c) Audio-Vision Cosine Similarity $\theta = 0.408$.



(d) Audio-Vision Cosine Similarity $\theta = 0.404$.



(e) Audio-Vision Cosine Similarity $\theta = 0.392$.



(f) Audio-Vision Cosine Similarity $\theta = 0.335$.

Figure 1: Audio-video consistency samples in AVSET.

174 with their corresponding confidence scores. Researchers can use these pseudo-labels to freely
175 partition the dataset.

176 We provide comprehensive meta-information for each video clip, including the YoutubeID of the
177 video, timestamps for each clip, audio-visual cosine similarity, a flag indicating whether sound
178 separation is required, and relevant text labels. For AVSET-10M (w/o. AVSET-700K), captions and
179 pseudo-labels are included, while AVSET-700K features manual audio labels.

180 C.3 Download URL

181 Please visit <https://avset-10m.github.io> to get the AVSET-10M. **Privacy Notice:** If any video
182 clips in this dataset infringe upon your privacy, please contact us for their removal.

183 C.4 LICENSE

184 AVSET-10M is released under the [CC BY 4.0] license. Before using this dataset, please ensure that
185 you have read and understood the terms of the license.

186 D Limitation

187 Since most existing video datasets predominantly contain clips with speech audio, which limits the
188 amount of non-speech samples, we plan to utilize more diverse data sources in the future. This
189 strategy aims to enhance the diversity of sample types and enable us to develop a more balanced and
190 expansive dataset.

191 E Ethical Impact

192 This paper primarily focuses on proposing a large-scale audio-visual correspondence dataset, aimed
193 at expanding research possibilities in the audio-visual sector. This field includes technologies like
194 video dubbing, which can lead to audio forgery. However, it's crucial to note that such dubbing does
195 not result in severe identity forgery issues, unlike those caused by voice cloning technologies.

196 References

- 197 [1] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu.
198 Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in*
199 *Neural Information Processing Systems*, 36, 2024.
- 200 [2] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao,
201 Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m:
202 Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*,
203 2024.
- 204 [3] Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, and Taylor Berg-Kirkpatrick.
205 Clipsep: Learning text-queried sound separation with noisy unlabeled videos. *arXiv preprint*
206 *arXiv:2212.07065*, 2022.
- 207 [4] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna
208 Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the*
209 *ACM*, 64(12):86–92, 2021.
- 210 [5] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing
211 Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for
212 audio events. In *2017 IEEE international conference on acoustics, speech and signal processing*
213 *(ICASSP)*, pages 776–780. IEEE, 2017.

- 214 [6] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala,
215 Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In
216 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages
217 15180–15190, 2023.
- 218 [7] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui
219 Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with
220 prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages
221 13916–13932. PMLR, 2023.
- 222 [8] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley.
223 Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM*
224 *Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.
- 225 [9] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-
226 to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing*
227 *Systems*, 36, 2024.
- 228 [10] Roman Solovyev, Alexander Stempkovskiy, and Tatiana Habruseva. Benchmarks and leader-
229 boards for sound demixing tasks, 2023.
- 230 [11] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito. The 2018 signal separation evaluation
231 campaign. In *Latent Variable Analysis and Signal Separation: 14th International Conference,*
232 *LVA/ICA 2018, Guildford, UK, July 2–5, 2018, Proceedings 14*, pages 293–305. Springer, 2018.
- 233 [12] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang,
234 Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative
235 and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.
- 236 [13] Zehan Wang, Ziang Zhang, Xize Cheng, Rongjie Huang, Luping Liu, Zhenhui Ye, Haifeng
237 Huang, Yang Zhao, Tao Jin, Peng Gao, et al. Molecule-space: Free lunch in unified multimodal
238 space via knowledge fusion. *arXiv preprint arXiv:2405.04883*, 2024.