Supplementary Material for

ACAV100M: Automatic Curation of Large-Scale Datasets for Audio-Visual Video Representation Learning

Sangho Lee*, Jiwan Chung*, Youngjae Yu, Gunhee Kim
Seoul National University

Thomas Breuel, Gal Chechik
NVIDIA Research

Yale Song
Microsoft Research

https://acav100m.github.io

A. On the Diversity of Concepts in Sampled Clips

A.1. Histogram of Cluster IDs

To analyze the diversity of concepts contained in our curated dataset, we examine the histograms of cluster IDs from the chosen videos. Figure 1 shows audio and visual histograms obtained from either our curated subsets or randomly sampled subsets at varying scales (20K, 200K, and 2M). To obtain these, we cluster the features from the last layer of audio and visual feature extractors, respectively, and plot the histograms of cluster IDs. For the purpose of visualization we sort the cluster indices by the cluster size in a decreasing order (and thus the cluster IDs do not match between "Random" and "Ours" in each of the plots). The histograms from random subsets represent the natural distribution of the entire video population.

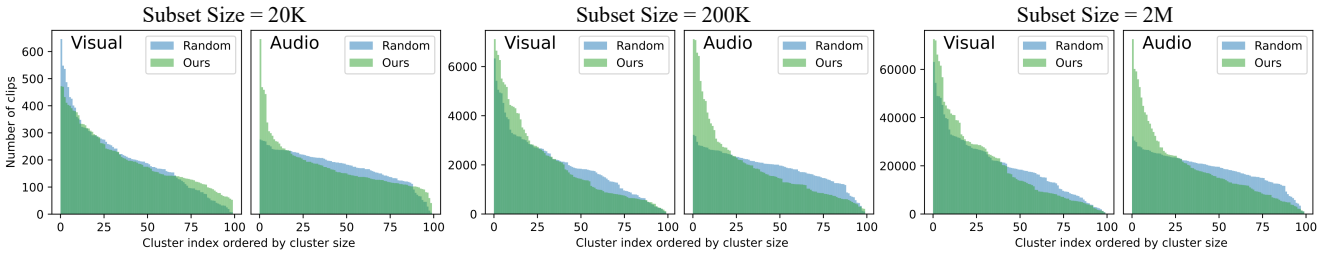


Figure 1. Histograms of cluster IDs from our curated subsets and randomly sampled subsets (with 100 cluster centroids). The blue histograms represent the case where samples are drawn uniformly random and thus is the unbiased representation of the concepts naturally appearing in the entire population.

In the visual domain, the curated datasets (green histograms) mostly follow the original cluster distributions (which is reflected in the blue histogram in each subplot). This indicates that the visual concept distribution largely follows the natural distribution in the entire population, suggesting that our subset contains visual concepts that are as diverse as the entire set.

On the other hand, the audio clusters show noticeable concentration in distribution after subset selection. Upon close inspection of videos from the largest audio clusters, we observe that our curated datasets tend to choose videos from clusters with high audio-visual correspondence (e.g., videos of a single person speaking with no other sound in background) while random sampling tend to choose videos from clusters with no apparent audio-visual correspondence (e.g., videos of multiple people taking with background music/noise). This shows that the concentration in the audio histograms is caused by filtering out videos of low audio-visual correspondence, which is a highly desirable artifact in the curated subset.

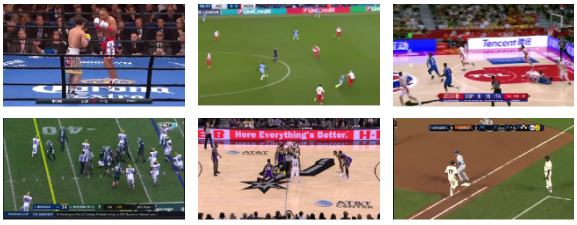
*Equal Contribution

Audio Clusters

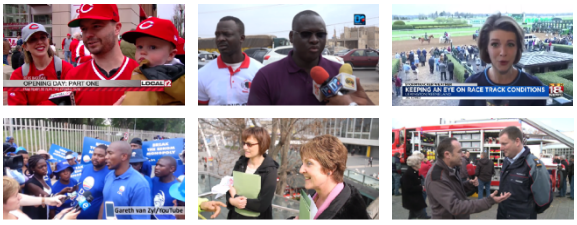
Cluster 91 (Size Ratio: 3.9%)
Audio: Female Voice, **Visual**: Woman Speaking

Cluster 44 (Size Ratio: 0.7%)
Audio: Metallic Sounds, **Visual:** Machine Parts, Tools

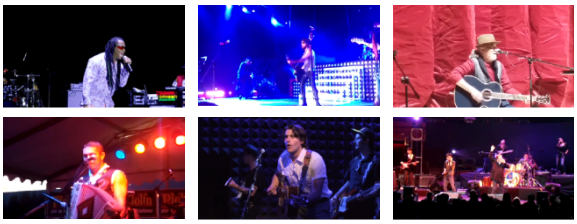
Cluster 89 (Size Ratio: 3.0%)
Audio: Commentaries, Crowd Cheering, **Visual:** Sports



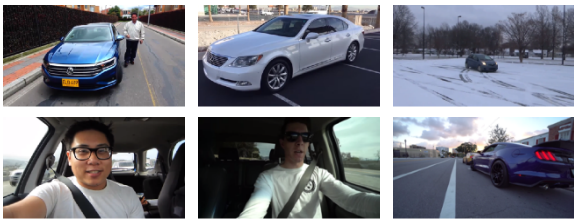
Cluster 37 (Size Ratio: 0.4%)
Audio: Voice, Background Noise, **Visual:** Outdoor Interview



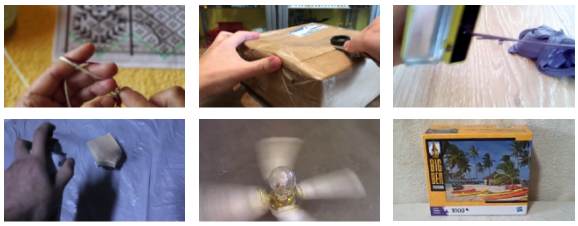
Cluster 67 (Size Ratio: 2.3%)
Audio: Singing, Crowd Cheering, **Visual:** Concert



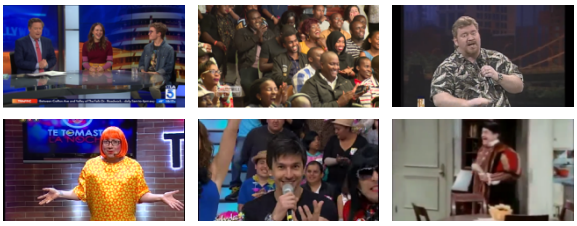
Cluster 33 (Size: 0.2%)
Audio: Engine Sound, **Visual:** Car



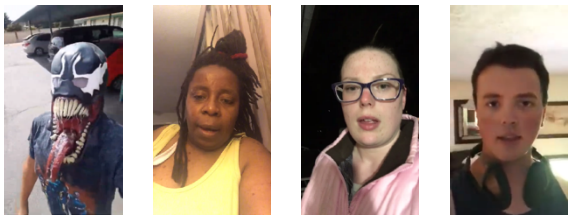
Cluster 51 (Size Ratio: 1.8%)
Audio: Object Sounds, **Visual:** Handling Objects



Cluster 46 (Size Ratio: 0.2%)
Audio: Laughing, Speech, **Visual:** Comedy



Cluster 77 (Size Ratio: 1.7%)
Audio: Phone Mic Recordings, **Visual:** Front Camera Selfies



Cluster 76 (Size Ratio: 0.1%)
Audio: Sizzling, Boiling, Stirring, **Visual:** Cooking

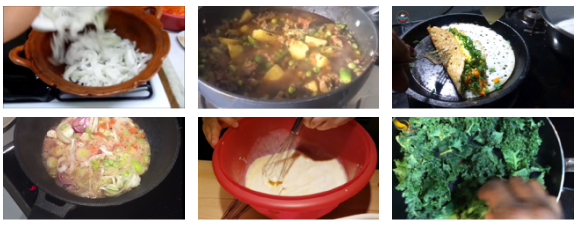
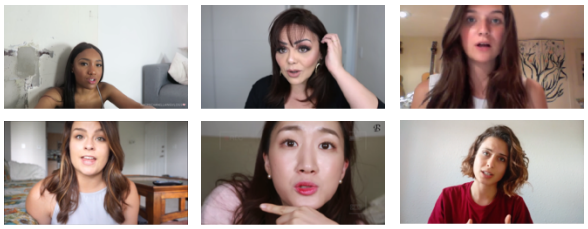


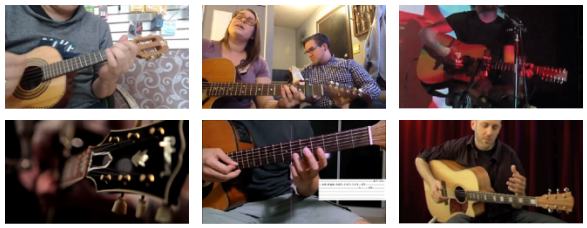
Figure 2. Representative samples and concepts derived from a manual inspection of 100 *audio* clusters of the 2M subset. We show samples from the five largest clusters on the left column and those from the five smallest clusters on the right. Each cluster captures distinctive audio-visual concepts, indicating that our curated subset contains various concepts with high audio-visual correspondence.

Visual Clusters

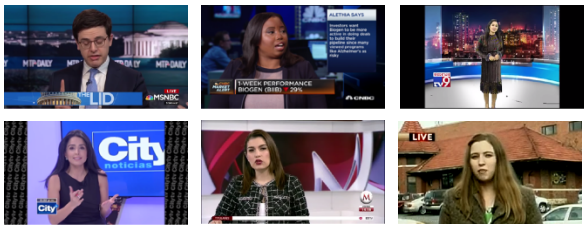
Cluster 83 (Size Ratio: 3.6%)
Audio: Female Voice, **Visual:** Woman Speaking



Cluster 0 (Size Ratio: 0.4%)
Audio: Guitar Sounds, Singing, **Visual:** Playing Guitar

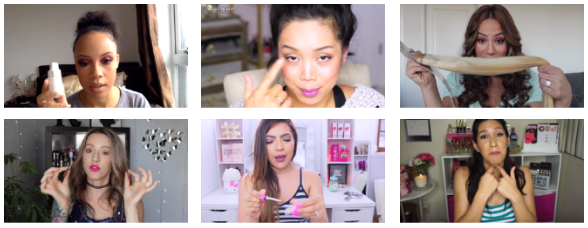


Cluster 42 (Size Ratio: 3.6%)
Audio: Clear Voice, **Visual:** News

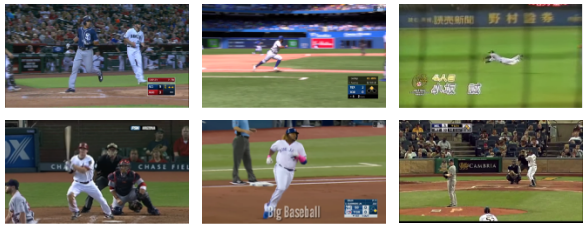


Cluster 2 (Size Ratio: 0.2%)
Audio: Punch, Crowd Noise, **Visual:** Martial Arts

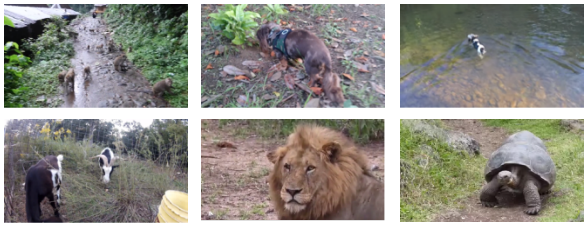
Cluster 33 (Size Ratio: 2.8%)
Audio: Brushing, Voice, **Visual:** Makeups



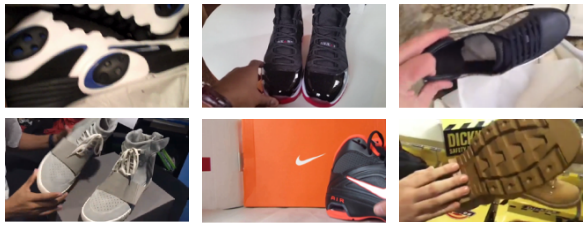
Cluster 76 (Size Ratio: 0.2%)
Audio: Hitting Balls, Crowd Noise, **Visual:** Baseball



Cluster 23 (Size Ratio: 2.2%)
Audio: Ambient Sounds, Animal Sounds, **Visual:** Nature



Cluster 88 (Size Ratio: 0.1%)
Audio: Object Sounds, **Visual:** Shoes Unboxing



Cluster 35 (Size Ratio: 1.9%)
Audio: Male Voice, **Visual:** Indoor Interview

Cluster 9 (Size Ratio: 0.1%)
Audio: Burning Sound, **Visual:** Fire

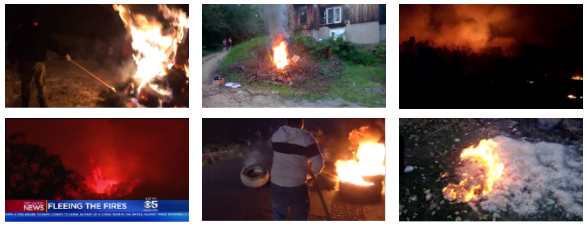


Figure 3. Representative samples and concepts derived from a manual inspection of 100 *visual* clusters of the 2M subset. We show samples from the five largest clusters on the left column and those from the five smallest clusters on the right. Each cluster captures distinctive audio-visual concepts, indicating that our curated subset contains various concepts with high audio-visual correspondence.

A.2. Qualitative Analysis of Audio-Visual Clustering Results

To further investigate the diversity of concepts appearing in our subsets, we manually inspect audio and visual clustering results in the 2M dataset and compare the concepts appearing in the largest clusters to those in the smallest ones. Figure 2 and Figure 3 show representative videos from the five largest and five smallest clusters obtained from audio and visual clustering results, respectively. Figure 2 (from audio clusters) suggests that our curated dataset contains diverse concepts including general sound categories (e.g., voice and objects sounds) as well as specific topics (e.g., outdoor interview and cooking). Similarly, Figure 3 (from visual clusters) also suggests that our dataset contains diverse concepts including both natural (e.g., animals and fire) and human sounds (e.g., makeup and playing guitar). Clips from larger clusters (depicted in the left column of Figure 2 and Figure 3) contain clear and isolated sound sources, while sounds of smaller clusters (the right column) are less distinguishable due to multiple sound sources or background noise. Our dataset also captures several audio-visual concepts that existing datasets (such as VGG-Sound [3] and AudioSet [6]) do not offer. For instance, in Figure 2, the 77th cluster contains videos recorded from a front-facing camera with voice recordings from a phone mic, and the 46th cluster contains videos of comedians performing exaggerated body actions with the sound of crowd (cheering and laughter). The 88th cluster in Figure 3 contains shoes unboxing videos.

B. Weighted Summation of Layer Scores (Section 4.2)

Table 1 compares different layer weighting schemes in clustering-based MI estimation, which shows that our multi-layer approach is generally robust to weight distributions. We explored two alternative weighting schemes: a *linear(k)* function with slope k and an *exp(k)* function with slope e^k; we used *uniform* weights in the main paper. We can see that precision is stable under a linear weighting scheme; the robustness comes from the `Combination` pairing approach which computes an average MI between all possible combinations across layers. However, precision drops significantly when the weights have a steep slope (e.g., *exp(-10)*), which is a degenerate case similar to the single-layer approach reported in Table 2 of the main paper.

Method	Layer Weights					Precision
	1	2	3	4	5	
exp(-10)	5e+10	2e+04	1	5e-05	2e-09	50.791
exp(-1)	7.4	2.7	1	0.4	0.1	65.374
exp(1)	0.1	0.4	1	2.7	7.4	79.858
exp(10)	2e-09	5e-05	1	2e+04	5e+10	57.880
linear(-0.50)	1.9	1.5	1	0.5	0.1	88.018
linear(-0.25)	1.5	1.2	1	0.8	0.5	88.673
linear(0.25)	0.5	0.8	1	1.2	1.5	88.777
linear(0.50)	0.1	0.5	1	1.5	1.9	87.997
Uniform (Ours)	1	1	1	1	1	88.705

Table 1. Different layer weighting schemes in clustering-based MI estimation using Kinetics-Sounds with `Combination` pairing.

C. Details of Linear Evaluation on Downstream Tasks (Section 5.1)

Table 2 shows the results of liner evaluation on downstream tasks, which were also shown in the bar chart of the main paper, Figure 4; we reproduced here to compensate for the lack of readability of the bar chart.

C.1. Experimental Settings

We pretrain audio-visual models in a contrastive manner [4] on different datasets. Specifically, we attach MLP projection heads on top of audio and visual feature extractors, respectively, and train the whole model end-to-end using the noise-contrastive loss (see Eqn. 1 of the main paper). As for the visual and audio backbone feature extractors, we use 3D ResNet-50 [2] and ResNet-50 [7], respectively. Each of the MLP projection head is composed of two fully-connected layers with ReLU [11] activation, and produces the embeddings of dimension 128. We pretrain the model for 50 epochs with a batch size 64. We use the AMSGrad variant [14] of AdamW [10] optimizer with a learning rate 1e-3, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and an L2 weight decay of 1e-5. We apply learning rate warm-up for the first 20,000 iterations followed by a linear decay of learning rate.

Size	Pretrain	UCF101		ESC-50		Kinetics-Sounds	
		top-1	top-5	top-1	top-5	top-1	top-5
-	Random Init	11.48	29.21	8.35	34.85	20.31	47.03
20K	Kinetics-Sounds	33.51	64.47	49.40	81.85	49.98	82.15
	Random Set	36.34	66.59	46.95	79.30	45.19	77.25
	Clustering (Ours)	46.28	75.24	50.55	81.30	55.78	85.15
200K	VGG-Sound	49.55	78.60	65.55	90.95	55.59	86.46
	Random Set	34.33	63.92	45.80	78.45	44.15	76.88
	Contrastive	45.10	76.46	56.90	85.00	53.80	85.26
	Clustering (Ours)	50.19	78.89	62.80	89.50	56.12	84.10
2M	AudioSet	55.54	83.94	65.05	90.70	57.46	86.72
	Random Set	41.12	72.24	52.75	83.55	48.30	79.54
	Contrastive	45.87	75.80	58.85	87.10	53.68	83.05
	Clustering (Ours)	55.63	83.92	65.10	90.50	57.48	87.19
10M	**Clustering (Ours)**	74.21	93.82	74.20	93.40	67.71	92.14
100M	**Clustering (Ours)**	86.10	97.94	86.95	97.45	75.42	95.88

Table 2. Linear evaluation of representations pretrained on different datasets. We report the top-1/5 accuracies (%) of video classification on UCF101 [15], audio classification on ESC-50 [13] and audio-visual classification on Kinetics-Sounds [1]. We average the accuracies across the official splits of UCF101 (three splits) and ESC-50 (five splits).

For linear evaluation on downstream tasks, we attach a linear classifier on top of the pretrained feature extractors and train it from scratch while fixing the parameters of the feature extractors. We use only the visual CNN for action recognition on UCF101 [15] and only the audio CNN for sound classification on ESC50 [13]. For audio-visual action recognition on Kinetics-Sounds [1], we concatenate audio-visual features before feeding them as input to the linear classifier. We apply dropout [8] with a 50% rate before the linear classifier. We train the model for 30 epochs with a batch size of 1024 on ESC-50 [13], for 10 epochs with a batch size of 64 on UCF101 [15] and for 5 epochs with a batch size of 64 on Kinetics-Sounds. We use the Adam [9] optimizer with a learning rate 1e-2, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and an L2 weight decay of 5e-6.

C.2. Impact of the Number of Centroids

To visualize the impact of the number of clusters in our clustering-based approach, we group the results by the number of clusters as shown in Figure 4. Notice that the number of clusters is not positively correlated with downstream task performance. Instead, clustering with about 500 clusters seems to yield the best performance. Also, experiments using the largest number of centroids ($C = 2000$) show low accuracy consistently across all datasets and subset sizes. This confirms our findings in Section 4.2 of the main paper: over-clustering tends to have a negative impact on the quality of the selected subset. We believe that this happens because, as the number of clusters increases, samples with homogeneous concepts in large clusters are scattered into small clusters sharing similar concepts. When we do not have many references to compare as in the early stage of subset selection, this fragmentation effect inhibits sample count sharing between conceptually similar small clusters, complicating the clustering-based MI estimation.

D. More Discussion on Subset Selection (Section 3.3.2)

D.1. Greedy Algorithm

We provide the details of the greedy algorithm [12] that is approximated using the batch greedy algorithm [5]. As shown in Algorithm 1, the greedy algorithm needs to re-evaluate the clustering-based MI estimator F on all the remaining candidates in each iteration. Thus, the time complexity is $O(N^2)$ where N is the size of the initial dataset $\mathbf{D}$.

On the other hand, the batch greedy algorithm approximates this by selecting the next element to be included in the solution within only a randomly chosen batch, not the entire candidates. This is shown in Algorithm 2 below (same as Algorithm 1 of the main paper; reproduced here for easy comparison).

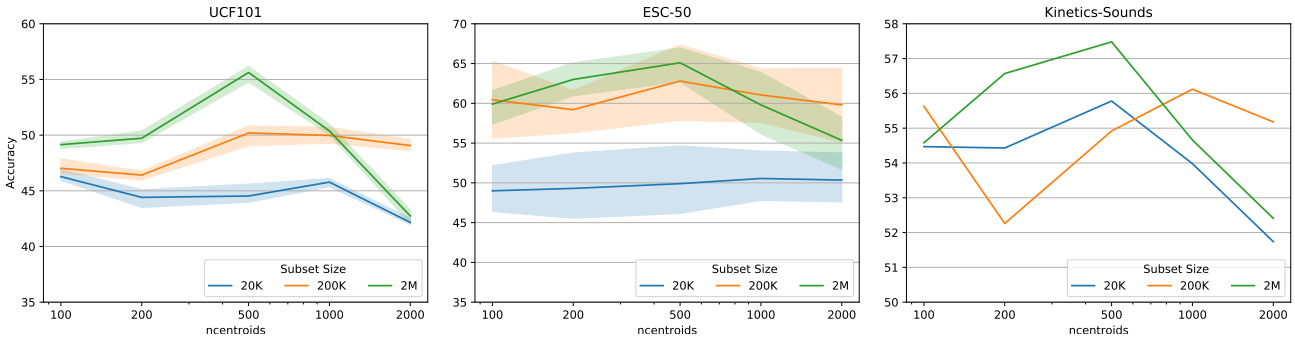


Figure 4. Linear evaluation of representations pretrained on the datasets that are constructed by our clustering-based approach. We report the top-1 accuracy (%) on UCF101 [15], ESC-50 [13], and Kinetics-Sounds [1], grouped by the number of cluster centroids. The shaded regions show 99% confidence intervals obtained by runs over the official splits of UCF101 (3 splits) and ESC-50 (5 splits).

Algorithm 1: Greedy Algorithm

Input: initial dataset $\mathbf{D}$, clustering-based MI estimator F, target subset size M
Output: $\mathbf{X} \subseteq \mathbf{D}, |\mathbf{X}| = M$
$\mathbf{X}_0 \leftarrow \emptyset$
for $i = 0$ **to** $M - 1$ **do**
  $x \leftarrow \text{argmax}_{x \in \mathbf{D} \setminus \mathbf{X}_i} F(\mathbf{X}_i \cup \{x\})$
  $\mathbf{X}_{i+1} \leftarrow \mathbf{X}_i \cup \{x\}$
end
$\mathbf{X} \leftarrow \mathbf{X}_M$
Return $\mathbf{X}$

Algorithm 2: Batch Greedy Algorithm (reproduced from the main paper for easy comparison)

Input: initial dataset $\mathbf{D}$, clustering-based MI estimator F, target subset size M, batch size b, selection size s
Output: $\mathbf{X} \subseteq \mathbf{D}, |\mathbf{X}| = M$
$\mathbf{X}_0 \leftarrow \emptyset, i \leftarrow 0$
while $|X_i| < M$ **do**
  Randomly sample $\mathbf{B} \subseteq \mathbf{D} \setminus \mathbf{X}_i, |\mathbf{B}| = b$
  $\mathbf{Y}_0 \leftarrow \emptyset, j \leftarrow 0$
  while $j < s$ **do**
    $x \leftarrow \text{argmax}_{x \in \mathbf{B} \setminus \mathbf{Y}_j} F(\mathbf{X}_i \cup \mathbf{Y}_j \cup \{x\})$
    $\mathbf{Y}_{j+1} \leftarrow \mathbf{Y}_j \cup \{x\}, j \leftarrow j + 1$
    if $|\mathbf{X}_i \cup \mathbf{Y}_j| = M$ **then** break
  end
  $\mathbf{X}_{i+1} \leftarrow \mathbf{X}_i \cup \mathbf{Y}_j, i \leftarrow i + 1$
end
$\mathbf{X} \leftarrow \mathbf{X}_i$
Return $\mathbf{X}$

D.2. Batch Greedy Subset Selection

When using the batch greedy algorithm for subset selection, the batch size b and the selection size s affect the quality of the selected subsets. We explore various (b, s) configurations on Kinetics-Sounds [1], as shown in Figure 5. Note that the performance gap between different batch sizes is small. The precision 93.9%, 94.3% and 94.6% are respectively obtained when using batch sizes $b = 40, 80, 160$ with the same ratio of selection size to batch size $s/b = 12.5\%$. On the contrary, the value of s/b highly affects the retrieval performance across all the batch sizes examined; the performance drops sharply as the ratio exceeds 25% regardless of the batch size. As stated in Section 4.2 of the main paper, we construct the dataset to have an equal number of positive and negative pairs and the drop in robustness manifests itself when the selection ratio s/b exceeds the *easy positive* ratio of 25%.

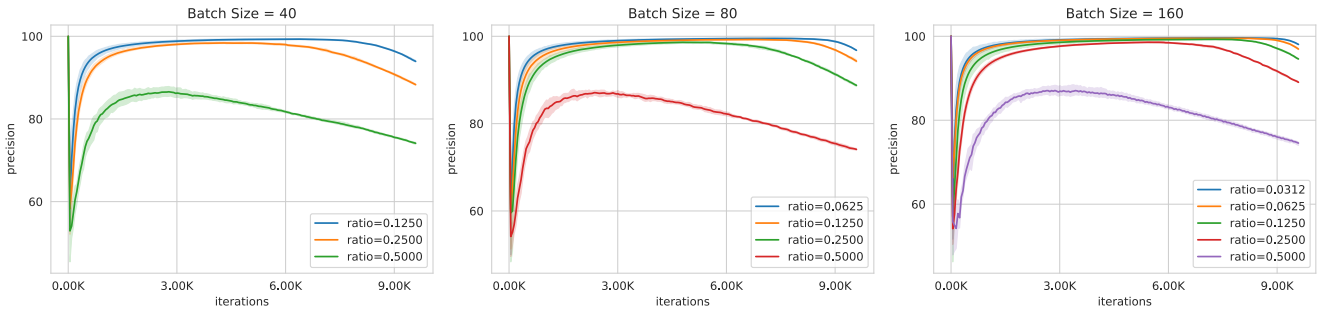


Figure 5. Precisions of Batch greedy algorithm with varying ratios of selection size to batch size, s/b (x axis: iterations, y axis: precision). We group the plots by the batch size: $b = 40, 80, 160$ from left to right. The shaded regions show 99% confidence intervals obtained by five runs on Kinetics-Sounds. The batch greedy algorithm is robust when the ratio is $\leqslant 25\%$, regardless of the batch size.

E. Details of Automatic Dataset Curation

Here, we describe the details of subset selection via (i) NCE-based MI estimation and (ii) clustering-based MI estimation. To construct datasets, we vary scales of 20K, 200K and 2M. Based on the results at the three scales, we also generate a version with 10M videos using the clustering-based approach.

E.1. NCE-Based MI Estimation

We use the linear projection heads that transform audio and visual features into 128-dimension embeddings. We randomly sample a subset of 100M clips from the initial 300M set that we crawl, and train on the subset for three epochs with a batch size $N_b = 1,024$. We use the AMSGrad variant of Adam optimizer [14] with a learning rate 2e-4, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We apply learning rate warm-up for the first 3 epochs followed by a linear decay of learning rate.

E.2. Clustering-Based MI Estimation

For SGD K-Means clustering, we train the cluster centroids with a mini-batch of size 100K for 100 epochs using a learning rate $\lambda =$ 1e-2. When applying the batch greedy algorithm, we use the fixed batch size $b = 10,000$ and the selection size $s = 500$ (with a ratio of $s/b = 0.05$), but vary the number of clusters $C \in \{100, 200, 500, 1000, 2000\}$ for each size of the datasets, except the dataset of 10M scale (we generate the dataset only with $C = 500$ for computational reasons).

F. Human Evaluation Interface (Section 5.2)

Figure 6 shows the user interface we developed for human evaluation. We provide guidelines on how to assess audio-visual correspondence:

> *You will watch a video clip for 10 seconds. Please determine whether there is audio-visual correspondence in the video. In other words, decide whether the sound source is visible or can be inferred from visual context.*

After a pilot study we gathered feedback from experts and added additional guidelines to help disambiguate common edge scenarios (shown in Figure 6). Annotators are given one 10-second clip at a time and asked to provide a Yes/No answer judging whether or not there is audio-visual correspondence in the given clip. We do not provide a replay interface to collect intuitive response from the raters.

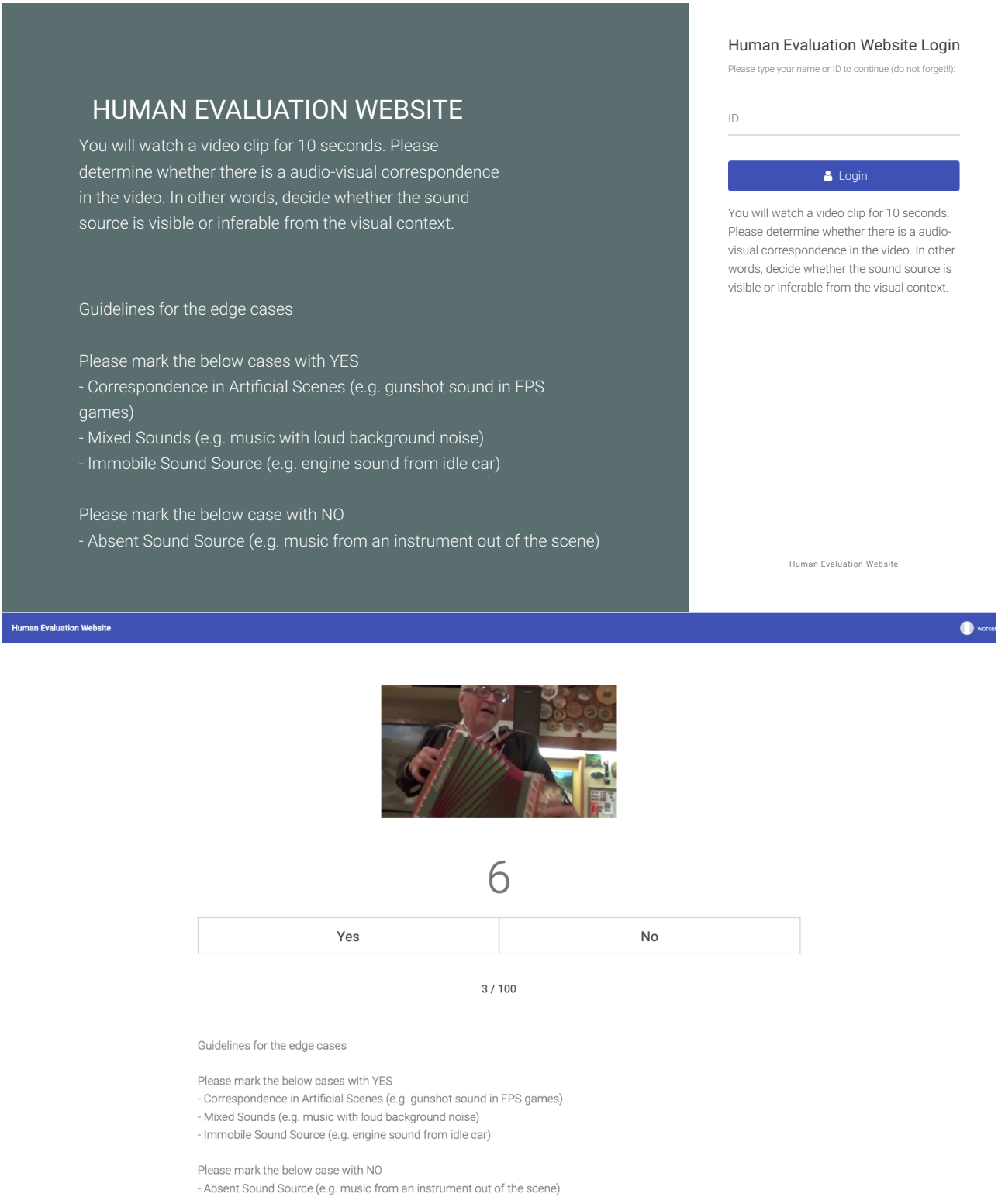


Figure 6. Screenshots of the human evaluation interface. The introduction page (top) provides instructions to the annotators, and the test page (bottom) shows clips to the raters and receives the corresponding Yes/No responses.

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, Listen and Learn. In *ICCV*, 2017.

[2] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, 2017.

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. VGGSound: A Large-Scale Audio-Visual Dataset. In *ICASSP*, 2020.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, 2020.

[5] Yuxin Chen and Andreas Krause. Near-optimal Batch Mode Active Learning and Adaptive Submodular Optimization. *ICML*, 2013.

[6] 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 *ICASSP*, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[8] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[9] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015.

[10] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.

[11] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *ICML*, 2010.

[12] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An Analysis of Approximations for Maximizing Submodular Set Functions—I. *Mathematical programming*, 14(1):265–294, 1978.

[13] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *ACM-MM*, 2015.

[14] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the Convergence of Adam and Beyond. In *ICLR*, 2018.

[15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402*, 2012.