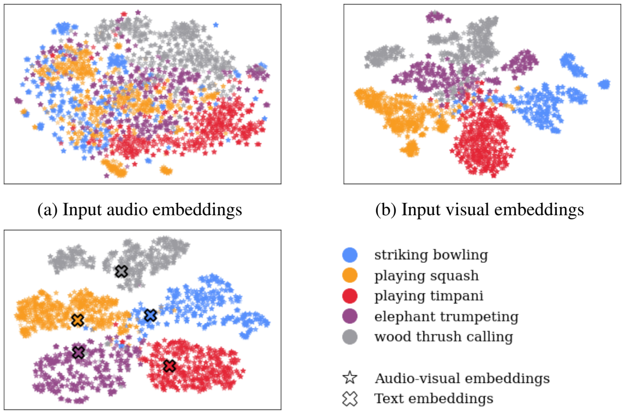
Supplementary Material: Audio-visual Generalised Zero-shot Learning with Cross-modal Attention and Language

In this supplementary material, we include additional qualitative results (Appendix A) and quantitative results (Appendix B) for our proposed audio-visual (G)ZSL framework.

A. Additional Qualitative Results

We provide additional qualitative results for our proposed AVCA model for the tasks of audio-visual GZSL and ZSL. We present t-SNE visualisations for the learnt audio-visual embeddings on the VGGSound-GZSL and UCF-GZSL datasets in Fig. 1 and Fig. 2.

In Fig. 1a, we can observe that the input audio features do not demonstrate a clear separation between the visualised classes for the VGGSound-GZSL dataset. The visual features exhibit a better clustering as can been seen in Fig. 1b. However, the visual features also include classes, such as *elephant trumpeting* and *wood thrush calling*, that are not clustered cleanly. Our AVCA model outputs multi-modal features that improve the clustering for both, seen and unseen classes (Fig. 1c). The learnt features for the two unseen classes *elephant trumpeting* and *wood thrush calling* are clustered and well-separated as opposed to the input features. This is impressive, since both classes were not included in the training set.

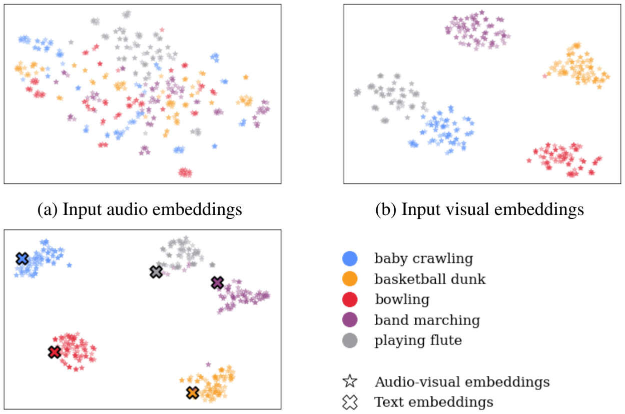


(a) Input audio embeddings

(b) Input visual embeddings

(c) Learnt audio-visual embeddings

Figure 1. t-SNE visualisation for three seen (*striking bowling, playing squash, playing timpani*) and two unseen (*elephant trumpeting, wood thrush calling*) test classes from the VGGSound-GZSL dataset, showing (a) audio and (b) visual features extracted with SeLaVi [4], and (c) learnt audio-visual embeddings of our model. Textual class label embeddings are visualised with a cross.



(a) Input audio embeddings

(b) Input visual embeddings

(c) Learnt audio-visual embeddings

Figure 2. t-SNE visualisation for three seen (*baby crawling, basketball dunk, bowling*) and two unseen (*band marching, playing flute*) test classes from the UCF-GZSL dataset, showing (a) audio and (b) visual features extracted with SeLaVi [4], and (c) learnt audio-visual embeddings of our model. Textual class label embeddings are visualised with a cross.

Similarly, for the UCF-GZSL dataset, we can observe in Fig. 2a that the input audio features are not grouped according to classes. In contrast, the visual input embeddings mostly exhibit a clear clustering of different classes. However, the classes *baby crawling* and *playing flute* are not well-separated as can be seen in Fig. 2b. This improves through learning, since the learnt audio-visual features in Fig. 2c show a clear divide between those two classes. In addition to that, the output embeddings for the unseen classes *band marching* and *playing flute* are overwhelmingly clustered well, too.

To summarise, our model learns to cluster both seen and unseen classes for different datasets by transferring information from the training data to unseen classes at test time.

Method type	Model	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
		S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
ZSL	ALE [2]	26.13	1.72	3.23	4.97	45.42	29.09	35.47	32.30	0.89	6.16	1.55	6.16
	SJE [3]	16.94	2.72	4.69	3.22	19.39	32.47	24.28	32.47	37.92	1.22	2.35	4.35
	DEVISE [5]	29.96	1.94	3.64	4.72	29.58	34.80	31.98	35.48	0.17	5.84	0.33	5.84
	APN [11]	6.46	6.13	6.29	6.50	13.54	28.44	18.35	29.69	3.79	3.39	3.58	3.97
Audio-visual ZSL	CJME [9]	10.86	2.22	3.68	3.72	33.89	24.82	28.65	29.01	10.75	5.55	7.32	6.29
	AVGZSLNet [8]	15.02	3.19	5.26	4.81	74.79	24.15	36.51	31.51	13.70	5.96	8.30	6.39
	AVCA	12.63	6.19	**8.31**	**6.91**	63.15	30.72	**41.34**	**37.72**	16.77	7.04	**9.92**	**7.58**

Table 1. Evaluating AVCA and state-of-the-art (G)ZSL methods for audio-visual GZSL and ZSL on the VGGSound, UCF, and ActivityNet (G)ZSL^{cls} benchmarks using features extracted from audio/video classification networks. We report the mean class accuracy on the seen (S) and unseen (U) test classes, and their harmonic mean (HM) for GZSL performance. The ZSL performance is evaluated on the test subset of samples from unseen classes.

B. Additional Quantitative Results

In this section, we provide additional quantitative results obtained with our AVCA. We present results for training and evaluating our AVCA model with a different set of input features in Appendix B.1. In particular, we use features extracted from networks that were pretrained for audio and video classification. We perform an additional ablation study that gradually transforms AVCA into AVGZSLNet [8] in Appendix B.2. Complete results that include the U and S performance for Table 3 in the main paper are provided in Appendix B.3. Finally, we give details about the number of parameters and GFLOPS required for training our AVCA model in Appendix B.4

B.1. Using features extracted audio/video classification networks

We additionally trained and tested our model and the baseline models using features extracted from audio and video classification networks (instead of the SeLaVi [4] features used in the main paper). In particular, the visual features were extracted with C3D [10], pretrained for video classification on Sports1M [7]. The audio features were extracted with VGGish [6], pretrained for audio classification on Youtube-8M [1]. We averaged the extracted features across time, resulting in a 4096-dimensional visual feature and a 128-dimensional audio feature for each video.

However, to use the audio features extracted from a network that was pretrained on Youtube-8M, we removed the test unseen classes from the VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL datasets that had an overlap with Youtube-8M. This resulted in slightly different dataset splits (VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls}) detailed in Table 2.

Dataset	# classes				# videos
	all	tr	v(U)	ts(U)	ts(U)
VGGSound-GZSL^{cls}	271	138	69	64	3200
UCF-GZSL^{cls}	48	30	12	6	845
ActivityNet-GZSL^{cls}	198	99	51	48	4052

Table 2. Statistics for our VGGSound, UCF, and ActivityNet (G)ZSL^{cls} datasets, showing the number (#) of classes in our splits (tr: train, v: validation, ts: test; S: seen, U: unseen). ^{cls} indicates the dataset splits that allow to use VGGish features pretrained on YouTube-8M. The full details about the dataset splits can be found at `https://github.com/ExplainableML/AVCA-GZSL`.

We provide results for training and evaluating our AVCA and the baselines using audio and video classification features in Table 1. AVCA outperforms all the baselines on all three datasets. On VGGSound-GZSL^{cls}, ACVA obtains a HM of 8.31% and ZSL of 6.91% compared to a HM of 6.29% for APN and a ZSL performance of 6.50% for APN. On UCF-GZSL^{cls}, AVCA obtains a HM of 41.34% and a ZSL of 37.72% compared to a HM of 36.51% for AVGZSLNet and a ZSL performance of 35.48% for DEVISE. On ActivityNet-GZSL^{cls}, AVCA outperforms AVGZSLNet with a HM of 9.92% compared to 8.30% and a ZSL of 7.58% compared to 6.39% for AVGZSLNet. These results show that AVCA outperforms the other competitors also when using audio and video classification features, proving again that our cross-attention mechanism and training objective provide a boost in performance.

B.2. Ablating AVCA in relation to AVGZSLNet

We additionally perform an ablation study that gradually transforms the AVCA model into AVGZSLNet [8] in Table 3. We show how our model components influence the (G)ZSL performance, resulting in our AVCA

Model	VGGSound-GZSL		UCF-GZSL		ActivityNet-GZSL	
	HM	ZSL	HM	ZLS	HM	ZSL
AVGZSLNet [8]	5.83	5.28	18.05	13.65	6.44	5.40
W/o x-att	6.02	4.81	26.82	18.37	6.50	5.64
W x-att with l_c loss	4.88	4.55	19.38	12.95	11.58	8.40
AVCA	**6.31**	**6.00**	**27.15**	**20.01**	**12.13**	**9.13**

Table 3. Ablation that gradually transforms our AVCA model into AVGZSLNet [8]. W/o x-att optimises each branch in isolation and their output predictions are averaged. x-att denotes cross-attention. l_c loss is the loss function used to train AVGZSLNet.

Model	VGGSound-GZSL			UCF-GZSL			ActivityNet-GZSL		
	S	U	HM	S	U	HM	S	U	HM
Visual branch	7.02	3.68	4.83	50.18	13.21	20.92	11.80	5.53	7.53
Audio branch	7.74	2.55	3.84	12.99	10.78	11.78	4.56	3.87	4.19
AVCA	14.90	4.00	**6.31**	51.53	18.43	**27.15**	24.86	8.02	**12.13**

Table 4. Influence of *training* AVCA with different modalities for GZSL on the VGGSound-GZSL, UCF-GZSL and ActivityNet-GZSL datasets measuring the GZSL performance on seen (S) and unseen (U) test classes and their harmonic mean (HM). Using both modalities yields the strongest GZSL performances.

model that outperforms AVGZSLNet on all three datasets. For this ablation, we use the SeLaVi [4] features and the same setup as in the main paper. W/o x-att corresponds to AVGZSLNet trained with our loss function (without our cross-attention). It can be observed that W/o x-att provides improvements on UCF-GZSL, with a HM of 26.82% compared to 18.05% and a ZSL performance of 18.37% compared to 13.65%. W x-att with l_c loss corresponds to AVGZSLNet with cross-attention and with the loss function proposed for AVGZSLNet. In this case, it can be observed that the cross-attention improves the results over AVGZSLNet with a HM of 11.58% compared to 6.44% and ZSL performance of 8.40% compared to 5.40% on ActivityNet-GZSL. These improvements can also be observed on the other datasets, showing that our novel loss and our cross-attention mechanism improve the performance over AVGZSLNet.

B.3. Extended results for training AVCA with different modalities

In this section, we extend the ablation study that uses different modalities for training (Table 3 in the main paper) by adding the performance on the seen (S) and unseen (U) test classes for all the datasets in Table 4.

On all three datasets it can be observed that there is an increase in both seen and unseen performance when using AVCA compared to using the Visual branch or the Audio branch. On VGGSound-GZSL, we can observe that the S performance for AVCA is 14.90% compared to 7.74% for the Visual branch. The U performance on VGGSound-GZSL is also stronger for AVCA than for the Visual branch, with a score of 4.00% compared to 3.68%. On the UCF-GZSL dataset, the S performance increases only slightly, from 50.18% for the Visual branch to 51.53% for AVCA. However, there is a significant increase in the U performance, from 13.21% for the Visual branch to 18.43% for AVCA. Finally, on ActivityNet-GZSL, AVCA yields a S score of 24.86% compared to 11.80% for the Visual branch. The U performance increases from 5.53% for the Visual branch to 8.02% for AVCA. These results show that the S/U performance increases significantly when using AVCA compared to the Visual branch or the Audio branch, leading to better HM/ZSL performances.

B.4. Number of parameters in AVCA.

AVCA contains 1.69M parameters in total, which is comparable to the 1.32M parameters used in AVGZSLNet [8]. ALE/SJE/DEVISE are significantly smaller with only 307.2k parameters. AVCA has a computational complexity of 2.36 GFLOPS, while AVGZSLNet has a computational complexity of 1.38 GFLOPS. Again, the fewest GFLOPS are required for ALE/SJE/DEVISE which have a computational complexity of 0.32 GFLOPS. These statistics show that AVCA is comparable to AVGZSLNet while providing significantly better results on all three datasets.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE TPAMI*, 2015.

[3] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

[4] Yuki M. Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, 2020.

[5] Andrea Frome, Greg Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2013.

[6] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017.

[7] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[8] Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. Avgzslnet: Audio-visual general-

ized zero-shot learning by reconstructing label features from multi-modal embeddings. In *WACV*, 2021.

[9] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *WACV*, 2020.

[10] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[11] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020.