UAVM: Towards Unifying Audio and Visual Models

Yuan Gong, *Member, IEEE,* Alexander H. Liu, Andrew Rouditchenko, and James Glass, *Fellow, IEEE*

***Abstract*—Conventional audio-visual models have independent audio and video branches. In this work, we *unify* the audio and visual branches by designing a Unified Audio-Visual Model (UAVM). The UAVM achieves a new state-of-the-art audio-visual event classification accuracy of 65.8% on VGGSound. More interestingly, we also find a few intriguing properties of UAVM that the modality-independent counterparts do not have.**

***Index Terms*—audio-visual learning, unified model**

I. Introduction

Humans perceive and understand their world by combining different sensory input modalities including sound and vision. To enable AI systems to have a similar ability, *audio-visual multi-modal learning* has been extensively studied [1], [2]. Due to the inherent differences between audio and video, conventional audio-visual learning methods typically either use handcrafted modality-specific features or have two *independent* audio and visual branches with different *architectures*, *training schemes*, and *model weights* [3], [4], [5], [6], [7].

This modality-specific paradigm began to change after the Transformer [8] showed its effectiveness for various tasks and modalities including audio [9], [10], [11] and video [12], [13]. Specifically, the Perceiver [14] model shows that it is possible to handle arbitrary configurations of different modalities using a *unified model architecture*, although the training method is still modality-specific. In addition to unified model architecture, data2vec [15] further demonstrates that a *unified training scheme* can be applied to different modalities to obtain state-of-the-art performance. Despite the unified model architecture and training scheme, the models of different modalities are trained independently and have independent weights. To go one step even further, SkillNet [16], EAO [17], VATT [18], and PolyViT [19] share part of the *model weights* among different modalities. Specifically, SkillNet [16] and EAO [17] are single models that handle multiple modalities, but different parts of the model weights are specialized for processing different modalities. VATT [18] uses a modality-agnostic, single-backbone Transformer and shares weights among different modalities. However, the training scheme of VATT is modality-asymmetric. PolyViT [19] shares model weights except the input tokenizer and task head.

The motivations for cross-modality model unification are multifold. First, it minimizes the use of handcrafted priors and inductive biases for each individual modality, which is more data-driven and saves manual effort. Second, a unified model that handles multiple modalities can be more parameter-efficient than a set of modality-specific models. Third, unified models are ideal *foundation models* [20] that can be adapted to a wide range of downstream tasks of multiple modalities.

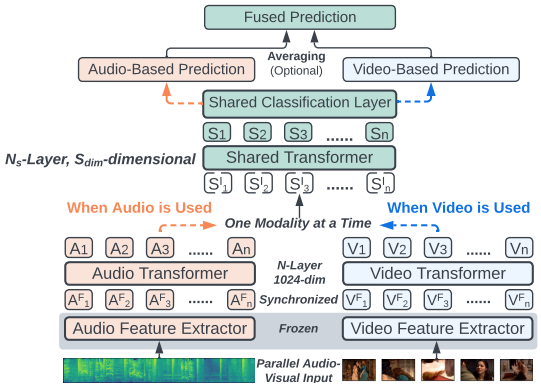

Fig. 1. An illustration of the unified audio-visual model (UAVM).

However, how unified models differ from modal-specific models remains unclear. Are unified models just two modal-independent models "glued" together? Does a unified model indeed encode audio and visual input into a unified representation space? We answer these questions by building and analyzing a Unified Audio-Visual Model (UAVM) that unifies 1) model architecture; 2) weights for high layers including the decision layer; and 3) training with a unified algorithm for audio and video. On VGGSound [21], UAVM achieves a new state-of-the-art accuracy for audio-visual event classification.

II. The Unified Audio-Visual Model

A. UAVM Model Architecture

The unique aspect of the UAVM is the shared Transformer and classification layer whose weights are shared between different modalities. This feature enables the UAVM to process both audio *and* video independently. Before the shared Transformer, we still have modal-specific feature extractors and optional modal-specific Transformers for each modality.

As shown in Figure 1, the audio or video is first input to the corresponding modality-specific feature extractor. Both audio and video feature extractors are ConvNeXt-Base [22]. For video, we use ImageNet pretrained ConvNeXt as the feature extractor. We uniformly sample RGB frames from the video at 3 FPS, input each frame to the ConvNeXt-Base, and get the mean-pooled penultimate layer representation of 1024-dimensions, i.e., for each 10-second video, the video features are 30 1024-dimensional vectors $\{V_1^F, ..., V_{30}^F\}$. For audio, we train a ConvNeXt using in-domain audio data (AudioSet or VGGSound) with ImageNet initialization [23] and then use it as the audio feature extractor. Each 10-second waveform is first converted to a 1000×128 log Mel filterbank (fbank) feature vector computed with a 25ms Hanning window every 10ms and input to the feature extractor. The output of the penultimate layer of ConvNeXt is a 30 (time)$\times$4 (frequency)$\times$1024 tensor.

Code at github.com/yuangongnd/uavm. The authors are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA (e-mail: {yuangong, glass}@mit.edu).

Algorithm 1 UAVM Model Training and Inference

Require: Dataset $\mathcal{D} = \{A, V, Label\}$, UAVM Model $\mathcal{M} = \{\theta_a, \theta_v, \theta_s\}$

Training ($\mathcal{D}, \mathcal{M}, \lambda_{\text{MT}}$)

while $i < max\ training\ iteration$ **do**
  ▷ One modality is used to train the model at an iteration
  if $unif(0,1) <$ modality training weight λ_{MT} **then**
    sample a batch of audio $\{A_i, Label_i\}$
    $Pred_a = \mathcal{M}_{\{\theta_a,\theta_s\}}(A_i)$
    $\mathcal{L} = Loss(Label, Pred_a)$
    backpropagate and update $\{\theta_a, \theta_s\}$
  else
    sample a batch of video $\{V_i, Label_i\}$
    $Pred_v = \mathcal{M}_{\{\theta_v,\theta_s\}}(V_i)$
    $\mathcal{L} = Loss(Label, Pred_v)$
    backpropagate and update $\{\theta_v, \theta_s\}$
return $\mathcal{M}$

Inference ($\{A, V\}, \mathcal{M}$)

if $A\ != None$ and $V\ != None$ **then**
  $Pred = (\mathcal{M}_{\{\theta_a,\theta_s\}}(A) + \mathcal{M}_{\{\theta_v,\theta_s\}}(V))/2$
else if one modality is missing **then**
  $Pred = \mathcal{M}_{\{\theta_v,\theta_s\}}(V)$ when $A == None$
  $Pred = \mathcal{M}_{\{\theta_a,\theta_s\}}(A)$ when $V == None$
return $Pred$

We apply a frequency mean pooling to produce 30 1024-dimension audio features $\{A_1^F, ..., A_{30}^F\}$. Note that we intend to make the audio and video features synchronized. Both feature extractors are frozen during training.

We then input the L2-normalized $\{A^F\}$ or $\{V^F\}$ to corresponding N-layer modality-specific Transformers to produce $\{A_1, ..., A_{30}\}$ or $\{V_1, ..., V_{30}\}$. After that, either $\{A_1, ..., A_{30}\}$ or $\{V_1, ..., V_{30}\}$ is used as an input $\{S_1^I, ..., S_{30}^I\}$ to the shared, modality-agnostic Transformer of N_s layers. This approach is fundamentally different from concatenating audio and visual tokens as input to the shared Transformer as in MBT [24] and Merlot Reserve [25]. We mean-pool the output of the shared Transformer $\{S_1, ..., S_{30}\}$ and input it to a shared linear classification layer. When only one modality is input, we use the output of the shared linear classification layer as the single-modality prediction; when both modalities are input, we run one forward pass for each modality and average the outputs of the two passes as a fused prediction (Algorithm 1). In other words, UAVM is robust to missing modalities.

Throughout the experiments, all Transformer layers have 4 attention heads and the modality-specific Transformer always has an embedding dimension of 1024. We tune the embedding dimension of the shared Transformer S_{dim} from 16 to 1024 to control its capacity. We fix the total number of Transformer layers to 6 ($N + N_s = 6$), and tune the number of modal-specific Transformer layers, N, and shared Transformer layers, N_s, from 0 to 6 to control the model unification level. When $N = 0$ and $N_s = 6$, the audio and visual models are maximally unified; when $N = 6$ and $N_s = 0$, the audio and visual models are completely independent including the classification layer. We set the model with $N = N_s = 3$ and $S_{dim} = 1024$ as the base UAVM model.

B. UAVM Model Training

We train UAVM with Algorithm 1. In each training iteration, only one modality is used. This is implemented by using a modality training weight λ_{MT} and uniform sampling, i.e., audio and video have λ_{MT} and $1 - \lambda_{\text{MT}}$ probability to be used, respectively. In other words, UAVM does not explicitly leverage the audio-visual correspondence information and can be trained with unpaired audio and video data (though in this paper, we focus on parallel datasets for simplicity). By default, we use modality training weight $\lambda_{\text{MT}} = 0.5$. As in prior work [9], [23], [24], we train the model with mixup [26], balanced sampling, label smoothing, and random time shifts. These training techniques are applied to both modalities with exactly the same hyperparameters, i.e., we use a unified training pipeline for the two modalities.

III. Experiment and Discussion

A. Experiment Settings

1) Dataset: We use two widely-used datasets for audio and video event classification: AudioSet [27] and VGGSound [21]. AudioSet is a collection of 2M 10-second YouTube video clips labeled with the sounds that the clip contains from a set of 527 labels. We downloaded 1.8M training and 17K evaluation audio and video samples for our experiment. VGGSound [21] is a collection of 200K 10-second YouTube video clips annotated with 309 classes. We download 184K training and 15K test samples for our experiments. One advantage of VGGSound is that the sound source is always visually present in the video clip. Also, its moderate size allows us to conduct extensive experiments with our computational resources. Therefore, while we compare UAVM performance with existing methods on both AudioSet and VGGSound, we conduct all ablation studies and analyses on VGGSound only.

2) Training Details: For all experiments, we train the model with a batch size of 144 and the Adam optimizer [28]. For the main experiments, we use an initial learning rate of 1e-5 and 5e-5 for AudioSet and VGGSound, respectively, and decrease the learning rate with a factor of 0.5 every epoch. We train the model for 10 epochs and report the last epoch result. For ablation studies, we tune hyper-parameters to ensure a fair comparison. We repeat all experiments 3 times with different random seeds and report the mean and standard deviation. The standard deviation is shown as shaded areas in Figure 2-5.

B. Model Performance Comparison

We compare the performance of three base models:

1. UAVM. The base UAVM described in Section II-A with 3 modal-specific Transformer layers and 3 shared Transformer layers ($N = N_s = 3$) and $S_{dim} = 1024$.

2. Modal-Independent Model. The base UAVM without shared Transformer layers (i.e., $N = 6, N_s = 0$), so the audio and visual models are completely independent including the classification layer.

3. Cross-Modal Attention Model. The base UAVM model with 3 modal-specific Transformer layers and 3 shared Transformer layers ($N = N_s = 3$) but instead of inputting one modality at a time to the shared Transformer, this model *concatenates* the outputs of modal-specific Transformers $\{A\}$ and $\{V\}$ together and inputs $\{A, V\}$ to the shared Transformer. This model only works when both modalities are input.

We show the results in Table I. Key conclusions are as follows: First, compared with the modal-independent model,

TABLE I
MODEL PERFORMANCE COMPARISON ON VGGSOUND AND AUDIOSET. (* SINGLE-MODAL MODEL TRAINED INDEPENDENTLY. † MODALITY-MISSING RESULTS OF A MULTI-MODAL MODEL.)

VGGSound (Top-1 Accuracy, %)	Audio	Video	Fusion
Chen *et al.* [21]	48.8	-	-
AudioSlowFast [29]	50.1	-	-
MBT [24]	52.3*	51.2*	64.1
Our Cross-Modal Attention Model	-	-	62.9±0.2
Our Modal-Independent Model	56.5±0.1*	49.7±0.2*	65.7±0.2
Our UAVM Model	**56.5±0.1**†	**49.9±0.2**†	**65.8±0.1**
Full AudioSet (mAP)	Audio	Video	Fusion
GBlend [30]	32.4*	18.8*	41.8
Attn Audio-Visual [31]	38.4*	25.7*	46.2
Perceiver [14]	38.4*	25.8*	44.2
MBT [24] (w/ 500k training samples)	44.3*	32.3*	52.1
Our Cross-Modal Attention Model	-	-	**50.4±0.1**
Our Modal-Independent Model	45.5±0.0*	26.8±0.1*	48.1±0.1
Our UAVM Model	**45.6±0.0**†	**27.4±0.1**†	48.0±0.0

UAVM achieves almost the same fusion performance when both modalities are input, and even slightly better results when a single modality is input with 76% of the parameters, demonstrating the feasibility of using a single network for two different modalities. Second, comparing UAVM with the "MBT-style" cross-modal attention model, we find a performance discrepancy between the datasets, i.e., UAVM is noticeably better on VGGSound while the cross-modal attention model is noticeably better on AudioSet. This is potentially because the sound source is always visually present in VGGSound videos but not in AudioSet videos. The UAVM strategy of giving equal weight to both modalities performs better than cross-modal attention models that could be dominated by one modality on tasks like VGGSound (video is always informative), but worse on AudioSet (video is not always informative). Finally, Transformer models with pretrained frozen features are a strong baseline with low computational cost. As a consequence, UAVM achieves a new SOTA performance on VGGSound, and outperforms all previous methods except MBT [24] on AudioSet. Note this is a fair comparison as both UAVM and MBT use ImageNet pretraining. Compared with MBT raw patches, the frozen feature input sequence length is much shorter (30 vs 1,500+), making it more computationally efficient since the Transformer has quadratic complexity.

We then conduct a series of ablation studies. We set models with $N = N_s = 3$ and $S_{dim} = 1024$ as the base UAVM and change one factor at a time to observe the performance change. First, we constrain the shared Transformer embedding dimension S_{dim} to smaller values to lower its capacity for UAVM and compare it with other models. For a fair comparison, the modal-independent and cross-modal attention models also have three 1024-dimensional Transformer layers and three S_{dim}-dimensional Transformer layers. With a limited capacity, the shared Transformer is forced to share neurons for two modalities. As shown in Figure 2 (upper), we find model performance generally improves with larger S_{dim}, but even when S_{dim} is very small, UAVM still performs similarly or even better than the modal-independent model. Second, we tune the number of shared layers N_s to change the level of model unification. Note that we fix the total number of Transformer layers to be 6, i.e., $N+N_s = 6$. As shown in Figure 2 (middle),

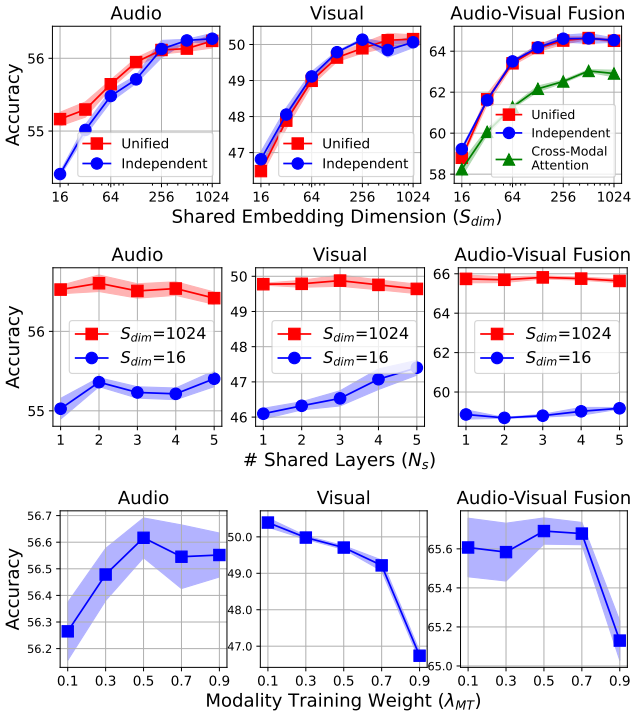

Fig. 2. The audio-based, video-based, and fused accuracy on VGGSound with various shared embedding dimension S_{dim} (upper), number of shared layers N_s (middle), and modality training weight λ_{MT} (lower).

when S_{dim} is small, the single-modality accuracy improves with more shared layers while when S_{dim} is large, the number of shared layers does not impact the performance much. This is true even when all 6 layers are shared (65.6% accuracy). Finally, as shown in Figure 2 (lower), even though audio and video are not equally informative for the event classification task (reflected in different accuracies), using a 0.5 λ_{MT} leads to optimal fusion results.

C. Unified Audio-Visual Representation Space

One core question about the unified model is if it indeed encodes two modalities in a unified latent space, or just processes each modality with a part of its parameters. We explore this by checking if a logistic regression model can successfully classify the input modality based on the mean-pooled penultimate layer representation of the UAVM model. Specifically, we use half of the VGGSound test set to train a logistic regression model and use the other half for testing. As shown in Figure 3.A, when the shared Transformer has a small S_{dim} and limited capacity, a logistic regression model is unable to predict the input modality based on the penultimate layer representation, in other words, the two modalities are encoded in a unified space. However, the modality classification accuracy gradually increases with S_{dim}, indicating that the model, though with shared weights, tends to encode two modalities in separate spaces when it has redundant capacity. Nevertheless, we do not see one or a small number of dimensions of the representation specifically encoding the input modality information because the classifier does not have a dominant coefficient (Figure 3.C). Further, as shown in Figure 3.B, the modality of the input and modal-specific layer (layer 1-3) representations can be perfectly classified but the modality classification accuracy drops suddenly with the first

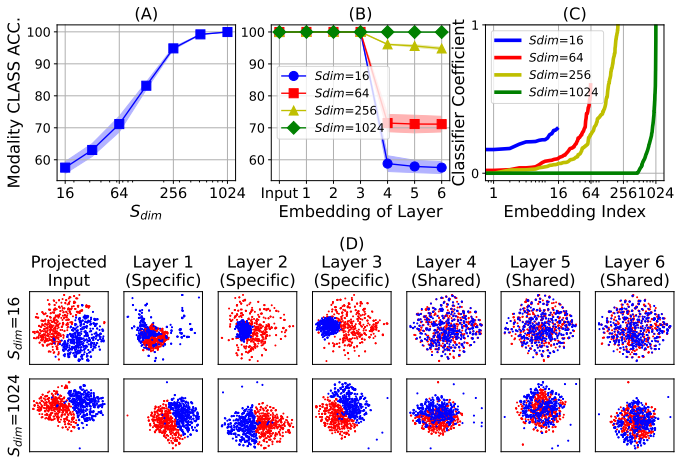

Fig. 3. (A) Modality classification accuracy based on the last layer representation of UAVM with various S_{dim}. (B) Modality classification accuracy based on intermediate representations of each layer of UAVM models. (C) The sorted modality classifier coefficients. (D) t-SNE plot of the intermediate representations of each layer with audio input (red) and video input (blue).

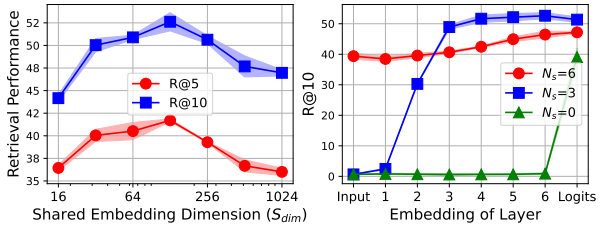

Fig. 4. (Left) A-V retrieval performance based on the last shared Transformer layer representation of UAVM models ($N_s = 3$) with various shared embedding dimensions S_{dim}. (Right) A-V retrieval performance based on the representation of various layers of models with $N_s = 0$ (fully independent model), $N_s = 3$, and $N_s = 6$ (fully unified model) and $S_{dim} = 128$.

shared layer (layer 4) representation, indicating that it is the shared Transformer that maps the two very different inputs to a unified space. The above findings can also be confirmed with the t-SNE plot of the representations of each layer with audio and video input shown in Figure 3.D.

D. Audio-Visual Representation Correspondence

One further question is how UAVM *aligns* paired audio and visual input in the latent representation space. Interestingly, we find a clear discrepancy between UAVM and modal-independent models. As a probing task, we calculate the A-V retrieval recall based on cosine similarity on a 1.5k subset of the VGGSound evaluation set (309 classes, 5 samples per class). Note that with reasonable single-modal classification accuracies, the audio and video of the same class can be naturally retrieved from each other. However, as shown in Figure 4 (right), for the modal-independent model ($N_s = 0$), the A-V retrieval recall is close to 0 for representations of all layers except the final linear classification head (i.e., the prediction logits) while for UAVMs, the A-V retrieval recall is much higher for representations for front layers (the more shared layers, the earlier a high A-V retrieval recall is achieved), indicating that the shared Transformer layers tend to propagate the supervision signal to front layers. Nonetheless, the unified models learn A-V correspondences beyond just intra-class correspondence as the A-V retrieval recall of UAVMs ($S_{dim} = 128$) is about 20% higher than the modal-independent model, while their classification accuracies are

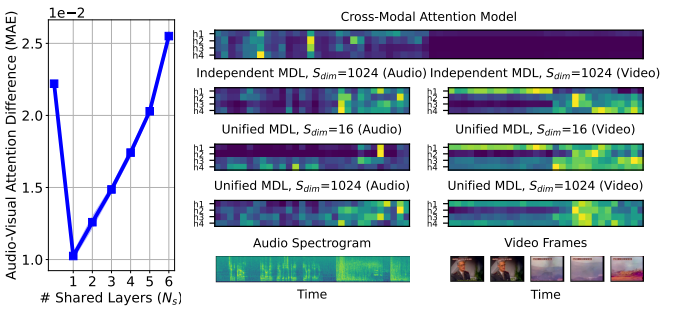

Fig. 5. (Left) Difference between the last layer attention maps with (paired) audio and video input for UAVM with various N_s. (Right) Temporal attention heatmap of 4 attention heads for a sample input (label: firing cannon).

similar. This is particularly interesting because UAVMs do not see simultaneous audio and video pairs, nor has any explicit audio-visual correspondence loss been applied during training. Further, as shown in Figure 4 (left), when $S_{dim} > 128$, the A-V retrieval recall starts to decrease with the increase of shared embedding dimension S_{dim}, while the classification accuracy consistently improves with larger S_{dim} (Figure 2, upper), indicating UAVM with redundant capacity is worse in audio-visual alignment and closer to modal-independent models, which is consistent with the finding in Section III-C.

E. Attention Maps

Even when the audio and video features are temporally synchronized, the informative part for event classification could be different. We show the temporal attention heatmap of 4 attention heads of the last Transformer layer in Figure 5 (right). The UAVM, even when the shared Transformer has very limited capacity ($S_{dim} = 16$), can pay attention to different parts of the audio and video, demonstrating its ability to process two modalities simultaneously. A cross-modal attention model, however, could be dominated by one modality while the other modality still contains useful information. Interestingly, by quantitatively calculating the mean absolute error (MAE) between attention maps of paired audio and video inputs, we find that the more modal-specific layers, the smaller the difference between the audio and video attention maps (Figure 5, left), indicating the modal-specific Transformers layers are forced to temporally align the informative part of the two modalities, so that succeeding shared Transformer layers can have a relatively more modal-agnostic attention map.

IV. Conclusion

In this work, we unify the audio and visual branches of a multi-modality model and build UAVM. Performance-wise, UAVM is similar to a modal-independent model, and outperforms cross-modal attention models on VGGSound. We find the capacity of the weight-sharing network greatly impacts the behavior of UAVM. When its capacity is constrained, UAVM shows some intriguing unique properties, e.g., it maps audio and video in a unified latent space, and better aligns paired audio and video in the latent space. Such unique properties gradually disappear with increasing model capacity, and UAVM behaves closer to modal-independent models.

V. Acknowledgments and Disclosure of Funding

This research is supported by the MIT-IBM Watson AI Lab.

REFERENCES

[1] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE signal processing magazine*, pp. 96–108, 2017.

[2] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *International Journal of Automation and Computing*, pp. 351–376, 2021.

[3] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in neural information processing systems*, 2016.

[4] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *IEEE International Conference on Computer Vision*, 2017, pp. 609–617.

[5] A. Rouditchenko, A. Boggust, D. Harwath, B. Chen, D. Joshi, S. Thomas, K. Audhkhasi, H. Kuehne, R. Panda, R. Feris *et al.*, "Avlnet: Learning audio-visual language representations from instructional videos," in *Interspeech*, 2021.

[6] M. Monfort, S. Jin, A. Liu, D. Harwath, R. Feris, J. Glass, and A. Oliva, "Spoken moments: Learning joint audio-visual representations from video descriptions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 871–14 881.

[7] O. Chang, O. Braga, H. Liao, D. Serdyuk, and O. Siohan, "On robustness to missing video for audiovisual speech recognition," *Transactions on Machine Learning Research*, 2022.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[9] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Interspeech*, 2021, pp. 571–575.

[10] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-Supervised Audio Spectrogram Transformer," in *AAAI Conference on Artificial Intelligence*, 2022, pp. 10 699–10 709.

[11] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang *et al.*, "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing," in *ACL*, 2022.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[13] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, 2021, pp. 10 347–10 357.

[14] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *International conference on machine learning*, 2021, pp. 4651–4664.

[15] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*, 2022, pp. 1298–1312.

[16] Y. Dai, D. Tang, L. Liu, M. Tan, C. Zhou, J. Wang, Z. Feng, F. Zhang, X. Hu, and S. Shi, "One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code," *arXiv preprint arXiv:2205.06126*, 2022.

[17] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. S. Feris, D. Harwath, J. Glass, and H. Kuehne, "Everything at once-multi-modal fusion transformer for video retrieval," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 020–20 029.

[18] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *Advances in Neural Information Processing Systems*, pp. 24 206–24 221, 2021.

[19] V. Likhosherstov, A. Arnab, K. Choromanski, M. Lucic, Y. Tay, A. Weller, and M. Dehghani, "Polyvit: Co-training vision transformers on images, videos and audio," *arXiv preprint arXiv:2111.12993*, 2021.

[20] R. Bommasani, D. A. Hudson, E. Adeli *et al.*, "On the opportunities and risks of foundation models," *ArXiv*, 2021. [Online]. Available: https://crfm.stanford.edu/assets/report.pdf

[21] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 721–725.

[22] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[23] Y. Gong, Y.-A. Chung, and J. Glass, "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[24] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," *Advances in Neural Information Processing Systems*, pp. 14 200–14 213, 2021.

[25] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi, "Merlot reserve: Neural script knowledge through vision and language and sound," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 375–16 387.

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[27] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[29] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Slow-fast auditory streams for audio recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 855–859.

[30] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 695–12 705.

[31] H. M. Fayek and A. Kumar, "Large scale audiovisual learning of sounds with weakly labeled data," in *International Joint Conferences on Artificial Intelligence*, 2021, pp. 558–565.