CAV-MAE Sync: Improving Contrastive Audio-Visual Mask Autoencoders via Fine-Grained Alignment

Edson Araujo[1*] Andrew Rouditchenko[2] Yuan Gong[2] Saurabhchand Bhati[2]
Samuel Thomas[3] Brian Kingsbury[3] Leonid Karlinsky[3] Rogerio Feris[3,4]
James R. Glass[2] Hilde Kuehne[1,4,5]

[1]Goethe University of Frankfurt, [2]MIT, [3]IBM Research, [4]MIT-IBM Watson AI Lab, [5]Tuebingen AI Center/University of Tuebingen

Abstract

Recent audio-visual learning methods often use global audio representations, limiting fine-grained temporal alignment with visual frames, and face conflicting optimization goals between reconstruction and cross-modal alignment. We propose CAV-MAE Sync, extending CAV-MAE [11] to address these issues. We enhance temporal granularity by aligning audio segments with video frames. We disentangle objectives using dedicated global tokens for contrastive loss and patch tokens for reconstruction. Finally, we add register tokens to improve spatial localization. CAV-MAE Sync achieves state-of-the-art results on AudioSet, VGGSound, and ADE20K Sound for zero-shot retrieval, classification, and localization. Code is available at https://github.com/edsonroteia/cav-mae-sync.

1. Introduction

Humans perceive the world in a multimodal way where especially auditory and visual perception are very closely connected. As a result, jointly learning the representations of both modalities has been a longstanding active research topic in multimodal learning [1–3, 6, 19, 21, 26]. Specifically audio-visual alignment has been tackled from multiple perspectives, with major works focusing on contrastive learning [5, 20, 24], but also exploring fusion-based methods [15, 17, 22, 25]. Recently, multitask formulations combine multiple learning objectives and have emerged as a promising direction for audio-visual representation learning. In particular, CAV-MAE [11] introduced a framework that jointly optimizes contrastive alignment between modalities and masked reconstruction within each modality. By leveraging both cross-modal and intra-modal learning signals, this approach has emerged as a foundational architecture that has inspired several follow-up works [12, 16, 18].

We argue that while these methods have achieved impressive results, they face two key limitations: they typically align video frames with global audio representations (e.g., matching 10 seconds of audio to a single frame), and they force both contrastive and reconstruction objectives to share a single representation, creating competing optimization goals. To address these issues, we propose CAV-MAE Sync, a simple yet effective extension of CAV-MAE that leverages natural temporal alignment between modalities while relaxing joint representation constraints. Our approach tackles three challenges: (1) addressing granularity mismatch by treating audio as a temporal sequence aligned with video frames (Figure 1), (2) resolving tension between competing objectives by introducing separate global tokens for each task, and (3) incorporating learnable register tokens

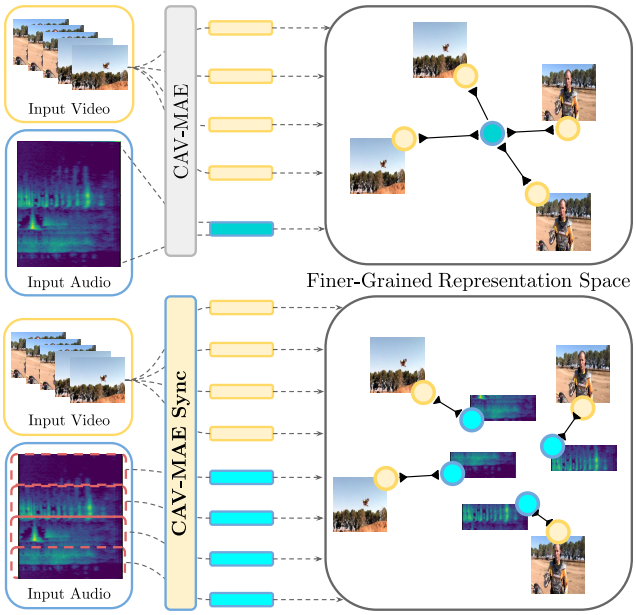

Figure 1. By representing audio with multiple finer-grained representations aligned with individual video frames, CAV-MAE Sync improves the precision of audio-visual alignment, in contrast to the original CAV-MAE, which uses a global audio representation that struggles with fine-grained temporal correspondence.

*araujo@em.uni-frankfurt.de

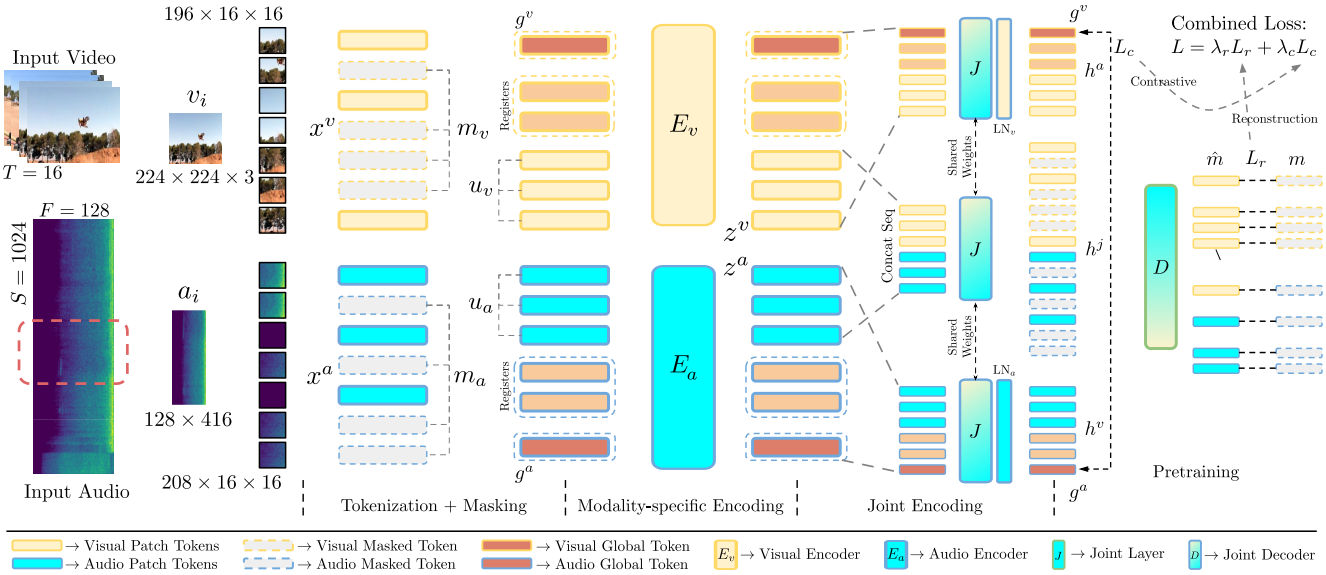

Figure 2. Overview of our approach. Our model processes video frames and audio segments in parallel through separate encoders E_a and E_v, with the audio encoder E_a operating on finer temporal granularity to better align with visual frames. Both modalities interact through the Joint Layer L and the Joint Decoder D The model is trained with both reconstruction and contrastive objectives.

that reduce semantic load on patch tokens while enabling finer-grained alignment. Experiments on VGGSound, AudioSet, and ADE20K datasets demonstrate that CAV-MAE Sync outperforms not only the original architecture but also competes with significantly more complex models across zero-shot retrieval, classification, and localization tasks.

Our contributions can be summarized as follows: (1) We propose CAV-MAE Sync, an extension of the CAV-MAE architecture that allows for a fine-grained temporal resolution on the audio side to support direct vision-audio alignment. (2) We introduce global tokens to disentangle the inhibiting contrastive and reconstruction objectives and add registers to the pipeline to further de-noise the ViT signal. (3) We evaluate the proposed setup on various downstream tasks and show a superior performance, even compared to significantly more complex architectures.

2. CAV-MAE Sync

2.1. Overview

Our method employs the contrastive masked autoencoder framework [11], training the model to reconstruct both visual and audio signals while enhancing audio-visual alignment through a contrastive objective. Unlike traditional approaches that utilize a single audio representation, we implement a sequence of audio representations temporally aligned with visual frames. This strategy ensures more coherent temporal alignment between audio and visual modalities without complicating the model architecture. For downstream tasks, we leverage the finer-grained audio-visual correspondences learned during pretraining. Figure 2 illustrates the data flow of our approach. In the following subsections, we first review the basics of CAV-MAE and then extend it in a second step toward the proposed CAV-MAE Sync framework.

2.2. Background: CAV-MAE

CAV-MAE processes video-audio pairs by sampling a frame v_i and using the full Mel spectrogram of the corresponding audio a_i. Both inputs are patchified, randomly masked, and tokenized into sequences u_v and u_a, incorporating positional and modality embeddings. These unmasked token sequences are passed through separate Vision Transformer (ViT) encoders, E_v and E_a, yielding modality-specific representations z^v and z_a. While sharing the same architecture and initialization, the encoders are trained independently. A joint transformer layer J processes these representations using three forward passes with shared weights but distinct layer normalizations: for visual tokens (z^v), for audio tokens (z_a), and one for their concatenation ($[z^v; z_a]$).

For the contrastive objective, global representations c_i^v and c_j^a are obtained by averaging the output patch tokens (h^v, h^a) from the single-modality passes. An InfoNCE loss L_c aligns these representations:

$$L_c = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(s_{i,i}/\tau\right)}{\sum_{k\neq i}\exp\left(s_{i,k}/\tau\right)+\exp\left(s_{i,i}/\tau\right)}\right) \tag{1}$$

where $s_{i,j} = \|c_i^v\|^T\|c_j^a\|$ is the cosine similarity between normalized representations, and τ is the temperature.

For the masked autoencoding (MAE) objective, the output from the joint (concatenated) pass is used to predict the original masked patches ($\hat{m}_i^a, \hat{m}_i^v$). The reconstruction loss terms L_i^a and L_i^v compute the Mean Squared Error (MSE) between predicted and original masked patches:

$$L_i^a = \frac{\sum_{i \in |m_a|} \left(\hat{m}_{a_i} - m_{a_i}\right)^2}{|m_a|} \quad L_i^v = \frac{\sum_{i \in |m_v|} \left(\hat{m}_{v_i} - m_{v_i}\right)^2}{|m_v|} \tag{2}$$

The total reconstruction loss L_r averages the batch terms:

$$L_r = \frac{1}{N} \sum_{i=1}^{N} (L_i^a + L_i^v) \tag{3}$$

The final learning objective $L = \lambda_c L_c + \lambda_r L_r$ balances contrastive alignment and masked reconstruction.

2.3. Improving Temporal Granularity

We argue that matching a full audio sequence to a single random frame creates a weak contrastive objective for two reasons: (1) frames from different scenes map to the same audio if they're from the same video, and (2) longer audio segments often contain multiple audio classes (e.g., in AudioSet), resulting in imprecise audio-visual correspondences. To address this, we increase temporal granularity by extracting frame-aligned audio segments, leveraging natural temporal alignment between modalities.

Temporal Alignment Process. For a video with T frames and audio spectrogram of length S, we extract a fixed-length spectrogram segment (s_{length}) for each frame i. We map frame indices to spectrogram positions using $s_{\text{center}_i} = \lfloor i \cdot S/T \rfloor$, then extract a centered window from $s_{\text{start}_i} = s_{\text{center}_i} - \lfloor s_{\text{length}}/2 \rfloor$ to $s_{\text{end}_i} = s_{\text{start}_i} + s_{\text{length}}$, adjusting boundaries for edge cases.

2.4. Disentangling Joint Modality Encoding

In the original architecture [11], patches are optimized for both contrastive and autoencoder objectives using a shallow joint layer, creating competing optimization goals. We propose strategies to disentangle these objectives, enhancing model performance.

Global Token Integration. Rather than aggregating patch tokens for global representations [9, 11], we introduce dedicated global tokens (g^a and g^v) for the contrastive objective. This separation allows patch tokens to focus on reconstruction while global tokens handle cross-modal alignment. Global tokens optimize via contrastive loss while all parameters update through both objectives.

Register Tokens. We incorporate register tokens to address the issue of high-norm patch tokens functioning as computation nodes rather than visual features [7]. This further maintains the separation between reconstruction (patch tokens) and contrastive objectives (global tokens), improving semantic capture and localization capabilities. These register tokens are appended to u_v and u_a and processed through the joint layer alongside global tokens.

Adaptation of the Joint Layer. With these additions, the joint layer processes modality-specific representations more effectively. The contrastive loss L_c now exclusively uses global tokens, computing similarity as $s_{i,j} = \|g_i^v\|^T \|g_j^a\|$. This ensures the contrastive objective operates on high-level representations while patch tokens handle reconstruction. By disentangling these objectives, each task is optimized independently, leading to improved performance across representation learning and downstream tasks.

2.5. Downstream Tasks

2.5.1. Cross-Modal Retrieval

For cross-modal retrieval, we leverage multiple temporal tokens to capture fine-grained relationships between modalities, rather than using single global tokens. We forward all frames and corresponding audio segments through their respective encoders and joint layer, obtaining global tokens g^v and g^a after passing through the joint layer J with layer normalizations LN_v and LN_a.

Similarity Calculation. For video-to-audio retrieval between query visual tokens $V_q = \{g_1^v, ..., g_T^v\}$ and target audio tokens $A_t = \{g_1^a, ..., g_T^a\}$, we construct a similarity matrix $S = V_q A_t^\top$. The final similarity score is computed by averaging the diagonal elements of S: Similarity Score $= \frac{1}{T} \sum_{t=1}^{T} s_{t,t}$. This diagonal-focused approach emphasizes temporally corresponding token pairs, promoting retrieval based on both semantic and temporal alignment. For a batch of videos, we compute similarity scores between all query-target pairs to construct a ranking matrix R, with higher scores indicating better matches.

2.5.2. Classification

For classification, we sample all frames and corresponding audio segments from each video, effectively increasing the batch size to $B \cdot T$. We obtain global tokens g^v and g^a for each temporal step t in video i, concatenate them to form a sequence C_i of length T, and prepend a learnable CLS token to create $C_i' = \{\text{CLS}, C_{i,1}, \ldots, C_{i,T}\}$. Our classification head f_{cls} (a two-layer transformer followed by a linear projection) produces predictions $\hat{y}_i = f_{\text{cls}}(C_i')$. We use binary cross-entropy loss for multi-class tasks (AudioSet) and standard cross-entropy for single-class tasks (VGGSound).

2.5.3. Sound-Prompted Semantic Segmentation

For sound-prompted semantic segmentation, we extract the global audio token g^a and all visual tokens h^v, then compute the cosine similarity between each h^v and g^a. This forms a similarity matrix L corresponding to the 14×14 patch grid, which we upscale to the original 224×224 frame resolution to create the predicted localization map.

	Baselines	AudioSet Eval Subset			VGGSound Eval Subset			AudioSet Eval Subset			VGGSound Eval Subset		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Visual → Audio / Audio → Visual	VAB-Encodec [23]	39.5	65.4	74.6	33.5	63.3	74.3	37.5	64.0	73.7	34.9	62.7	73.1
	CAV-MAE [11]	16.1	38.6	49.3	14.7	35.3	45.9	9.5	22.6	32.4	8.3	23.8	32.4
	CAV-MAE$^{\text{Scale+}}$ [11]	18.8	39.5	50.1	14.8	34.2	44.0	15.1	34.0	43.0	12.8	30.4	40.3
	LanguageBind [28]	6.4	20.2	28.3	10.3	30.1	39.7	4.4	15.0	22.5	6.5	22.7	33.5
	AVSiam [18]	19.7	–	–	19.0	–	–	17.6	–	–	20.4	–	–
	ImageBind [10]	22.1	43.2	52.6	21.6	43.4	52.9	20.8	42.6	51.6	20.7	43.2	53.4
	Ours	**35.2**	**58.3**	**67.6**	**27.9**	**51.7**	**61.8**	**27.9**	**52.4**	**62.2**	**23.2**	**46.2**	**58.1**

Table 1. Zero-shot retrieval results on AudioSet and VGGSound for Visual to Audio (V→A) and Audio to Visual (A→V) tasks. Our model achieves state-of-the-art zero-shot performance across all retrieval metrics (R@1, R@5, R@10) on both datasets, surpassing baselines like ImageBind and AVSiam. Fine-tuned VAB-Encodec scores are provided as an upper bound for comparison.

Baselines	Pretrain Dataset	AS20K↑	VGGSound↑
VAB-Encodec [23]	AS-2M + VGGS	33.3	57.6
CAV-MAE [11]	AS-2M	27.3	-
CAV-MAE$^{\text{Scale+}}$ [11]	AS-2M	28.5	47.7
CAV-MAE$^{\text{Scale++}}$ [11]	AS-2M	29.2	51.1
CAV-MAE$^{\text{Scale+++}}$ [11]	AS-2M	25.3	51.6
MaViL [16]	AS-2M	30	-
Ours	AS-2M	**30.5**	**52.7**

Table 2. Comparing audio-visual classification performance using linear probing. Numbers reported for AS20K are calculated using mAP (mean Average Precision) and VGGSound with accuracy.

Baselines	**mAP ↑**	**mIoU ↑**
DAVENet [14]	16.8	17.0
CAVMAE [11]	26.0† / 21.2	20.5† / 20.9
ImageBind [10]	18.3	19.1
Ours	**22.6**	**22.7**

Table 3. Sound-prompted semantic segmentation: Comparison of sound localization methods on ADE20K Sound dataset [13]. †Original reported by [13] / our reproduction.

3. Evaluation

3.1. Downstream Tasks and Results

We evaluate our model on three downstream tasks: cross-modal retrieval, classification, and sound-prompted segmentation. Using the same pretrained model without task-specific modifications, we compare against state-of-the-art methods on AudioSet [8], VGGSound [4], and ADE20K_Sound [27]. Our evaluation assesses both representation quality and the model's ability to establish fine-grained audio-visual correspondences.

Zero-shot Audio-Visual Retrieval. We evaluate bidirectional retrieval (Visual→Audio and Audio→Visual) using Recall@k metrics (k={1,5,10}) on AudioSet and VGGSound test sets, following the protocol from CAV-MAE [11] with cosine similarity for ranking. As shown in Table 1, our model achieves state-of-the-art performance in both retrieval directions compared to CAV-MAE [11], ImageBind [10], AVSiam [18], and VAB [23], confirming that our temporally consistent audio-visual correspondences and disentangled contrastive MAE objective significantly improve generalization to retrieval tasks.

Classification. We assess representation quality through linear probing on AudioSet and VGGSound, freezing the pretrained encoder and training only a classifier. We report mean Average Precision (mAP) for AudioSet's multi-label task and accuracy for VGGSound's single-label setting. As shown in Table 2, our model achieves 30.5 mAP on AudioSet and 52.7% accuracy on VGGSound, outperforming CAV-MAE variants and MaViL with AudioSet-2M pretraining, demonstrating strong classification capabilities alongside retrieval performance.

Sound-Prompted Image Segmentation. Following [13], we evaluate cross-modal localization on ADE20K_Sound, where an audio prompt must guide segmentation of corresponding image regions. Performance is measured via mean Average Precision (mAP) and mean Intersection over Union (mIoU). Table 3 shows our model achieves 22.7 mIoU, comparable to prior self-supervised methods like CAVMAE and ImageBind, while using the same backbone.

4. Conclusion

In this work, we introduced CAV-MAE Sync, an extension of the popular CAV-MAE framework that addresses key challenges in audio-visual learning by treating audio as a temporally aligned sequence, disentangling contrastive and reconstruction objectives through separate global tokens, and enhancing spatial localization with learnable register tokens. Our experiments demonstrate across AudioSet, VGGSound, and ADE20K that these architectural improvements offer a more effective and efficient approach to audio-visual representation learning that harmoniously aligns temporal and spatial aspects of audio and visual modalities.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 208–224. Springer, 2020. 1

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *IEEE International Conference on Computer Vision*, pages 609–617, 2017.

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems*, 29, 2016. 1

[4] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725, 2020. 4, 1

[5] Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7016–7025, 2021. 1

[6] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3884–3892, 2020. 1

[7] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 2

[8] Jort F Gemmeke, Daniel PW 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*, pages 776–780, 2017. 4, 1

[9] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16144–16154, 2023. 3

[10] 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*, pages 15180–15190, 2023. 4

[11] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3, 4

[12] Yuxin Guo, Siyang Sun, Shuailei Ma, Kecheng Zheng, Xiaoyi Bao, Shijie Ma, Wei Zou, and Yun Zheng. Cross-mae: Cross-modality masked autoencoders for region-aware audio-visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26721–26731, 2024. 1

[13] Mark Hamilton, Andrew Zisserman, John R Hershey, and William T Freeman. Separating the" chirp" from the" chat": Self-supervised visual grounding of sound and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13117–13127, 2024. 4

[14] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665, 2018. 4

[15] Chiori Hori, Takaaki Hori, Gordon Wichern, Jue Wang, Teng-Yok Lee, Anoop Cherian, and Tim K Marks. Multimodal attention for fusion of audio and spatiotemporal features for video description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2528–2531, 2018. 1

[16] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. Mavil: Masked audio-video learners. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 4

[17] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *International conference on learning representations*, 2020. 1

[18] Yan-Bo Lin and Gedas Bertasius. Siamese vision transformers are scalable audio-visual learners. In *ECCV*, 2024. 1, 4

[19] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309, 2023. 1

[20] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *International Conference on Learning Representations*, 2021. 1

[21] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. In *Interspeech*, 2021. 1

[22] Arda Senocak, Junsik Kim, Tae-Hyun Oh, Dingzeyu Li, and In So Kweon. Event-specific audio-visual fusion layers: A simple and new perspective on video understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2237–2247, 2023. 1

[23] Kun Su, Xiulong Liu, and Eli Shlizerman. From vision to audio and beyond: A unified model for audio-visual representation and generation. In *International Conference on Machine Learning*, pages 46804–46822, 2024. 4

[24] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6420–6429, 2023. 1

[25] Qinghao Ye, Xiyue Shen, Yuan Gao, Zirui Wang, Qi Bi, Ping Li, and Guang Yang. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7950–7959, 2021. 1

[26] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *European Conference on Computer Vision*, pages 570–586, 2018. 1

[27] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 4

[28] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024. 4