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 advances in audio-visual learning have shown promising results in learning representations across modalities. However, most approaches rely on global audio representations that fail to capture fine-grained temporal correspondences with visual frames. Additionally, existing methods often struggle with conflicting optimization objectives when trying to jointly learn reconstruction and cross-modal alignment. In this work, we propose CAV-MAE Sync as a simple yet effective extension of the original CAV-MAE [14] framework for self-supervised audio-visual learning. We address three key challenges: First, we tackle the granularity mismatch between modalities by treating audio as a temporal sequence aligned with video frames, rather than using global representations. Second, we resolve conflicting optimization goals by separating contrastive and reconstruction objectives through dedicated global tokens. Third, we improve spatial localization by introducing learnable register tokens that reduce the semantic load on patch tokens. We evaluate the proposed approach on AudioSet, VGG Sound, and the ADE20K Sound dataset on zero-shot retrieval, classification, and localization tasks demonstrating state-of-the-art performance and outperforming more complex architectures. Code is available at https://github.com/edsonroteia/cav-mae-sync.

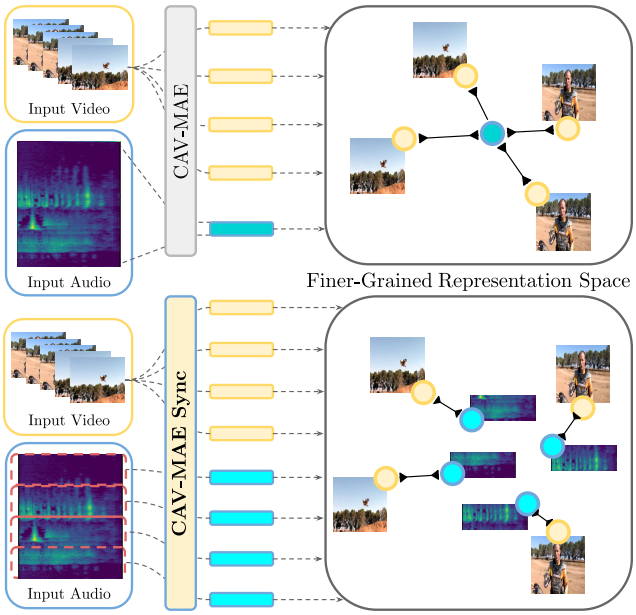

Figure 1. By representing audio with multiple finer-grained representations aligned with individual video frames, CAV-MAE Sync improves the precision of audio-visual alignment, in contrast to the original CAV-MAE, which uses a global audio representation that struggles with fine-grained temporal correspondence.

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, 2, 4, 9, 25, 34, 42]. Specifically audio-visual alignment has been tackled from multiple perspectives, with major works focusing on contrastive learning [8, 27, 38], but also exploring fusion-based methods [18, 23, 36, 40]. Recently, multitask formulations combine multiple learning objectives and have emerged as a promising direction for audio-visual representation learning. In particular, CAV-MAE [14] 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 [15, 20, 24].

We argue that while these methods have achieved impressive results, they show two significant limitations. First, most of them align the video and audio information based on a global audio representation, thus matching, for example, 10 seconds of audio to a single video frame. Second,

while the joint learning of cross-modal similarities together with in-modality reconstruction has proven to be a successful strategy, the vanilla implementation of this idea suffers from the fact that both objectives need to be achieved by a single representation, learned in a joint layer. This can be problematic as having a too similar representation for audio and vision can inhibit their reconstruction and vice versa.

To address this, we propose CAV-MAE Sync, a simple yet effective extension of the original CAV-MAE framework for audio-visual learning that takes advantage of the natural temporal alignment between modalities while at the same time relaxing the constraint of a joint representation. Namely, our work addresses three key challenges of this architecture as follows: *i*) First, we tackle the granularity mismatch between visual and audio modalities. While existing methods typically operate on global audio-visual alignment objectives, we propose treating audio as a temporal sequence of instances, better matching the inherent structure of both modalities. This process is illustrated in Figure 1. *ii*) Second, we resolve the tension between contrastive and reconstruction objectives. Current approaches like CAV-MAE simultaneously enforce representation similarity through contrastive loss while applying reconstruction loss on the same embedding layer. We argue this creates conflicting optimization goals. Our solution introduces separate global tokens, allowing each objective to operate in its optimal space. *iii*) Third, we introduce learnable register tokens into the pipeline. Similar to the results seen in [10], these registers further alleviate the semantic load of the patch tokens, while also allowing for finer-grained audio-visual alignment and thus better localization.

To evaluate our approach, we conduct experiments on zero-shot audio-visual retrieval, linear probing for audio-visual classification, and localization on the well-known VGGSound [6], AudioSet [11], and ADE20K [43] datasets. Our results show that the proposed CAV-MAE Sync is not only superior to the original CAV-MAE architecture but also competes with significantly more complex architectures and achieves state-of-the-art performance on all 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. Related Work

Contrastive Audio-Visual Masked Autoencoder Models. CAV-MAE [14] presented the first audio-visual model that leverages both contrastive learning and masked autoencoder objectives, and together with AVMAE [12], pioneered the self-supervised objective of masked autoencoding in the audio-visual domain. By combining the two learning tasks, CAV-MAE demonstrated strong performance across audio-visual tasks, effectively learning representations that capture both modality-specific features and cross-modal relationships. Building on CAV-MAE's success, several works have proposed improvements to this contrastive audio-visual masked autoencoder framework. CrossMAE [15] introduced a region-aware approach using dual encoders and a fusion module to predict masked regions in both modalities, enabling fine-grained cross-modal understanding. MaViL [20] built on top of the framework by performing video-level encoding and introducing a sophisticated masking strategy with both inter-modal contrasting between matched video-audio pairs and intra-modal contrasting between masked views of the same modality. AVSiam [24] proposed a parameter-efficient Siamese architecture that shares a single vision transformer backbone between modalities while maintaining the core CAV-MAE framework, reducing the model size while claiming to help bridge the modality gap between audio and visual representations. Unlike other CAV-MAE models, VAB [37] focuses on masked audio token prediction in latent space using pretrained tokenizers. While it can be fine-tuned with contrastive learning for retrieval tasks, its primary innovation is a visual-conditioned masked audio prediction that enables both representation learning and audio generation capabilities. While these methods have made significant advances, they mostly operate on global audio-visual alignment objectives, missing opportunities for finer temporal granularity. Our work addresses this limitation by treating audio as a temporal sequence of instances, introducing separate global tokens to help disentangle competing objectives, and enhancing spatial localization through register tokens, achieving state-of-the-art performance with a simpler architecture.

General Contrastive Audio-Visual Models. Initial approaches to audio-visual learning leveraged knowledge distillation [4, 31], where well-trained visual models guided the optimization of audio networks. This was followed by the emergence of paired sample discrimination methods [2, 3, 22, 30], which learned representations by distinguishing between matching and mismatched audio-visual pairs. Building on this successful paradigm, recent contrastive learning approaches [29, 34, 38] have formalized the learning objective by maximizing similarity between positive pairs while minimizing similarity between negative examples in the embedding space. Several works have focused on improving contrastive learning through better sampling and data augmentation strategies. This includes active learning approaches for mining hard negatives [26], robust sampling to handle temporal misalignment [28], and

multi-view techniques [32, 33, 39, 41] that leverage both global and local temporal context. Recent work has explored making representations more robust by relaxing temporal synchronicity constraints [35] and introducing equivariance [21]. In another direction, several works have explored localization capabilities emerging from contrastive learning. Audio-Visual Correspondence [3] demonstrated that localization naturally emerges from the audio-visual correspondence task without explicit supervision. Building on this, Audio-Visual Associative Localizations [17] developed methods to explicitly link spoken audio descriptions with specific image regions. Chen et al. [7] improved localization through hard negative mining in the contrastive learning process. Another line of work has further explored approaches to learning unified embeddings across multiple modalities. ImageBind [13] uses images as a semantic anchor, leveraging strong semantic relationships from image-text models to align multiple modalities like audio, depth, and thermal data. This enables zero-shot capabilities across modalities without requiring explicit pairings between all of them. Following, LanguageBind [44] argues for directly aligning modalities to language, utilizing a pre-trained language encoder as the semantic anchor point, demonstrating good performance on language-related tasks through this direct alignment strategy. More recently, DenseAV [16] achieved strong localization by aligning dense audio-visual features through a dual encoder architecture built on top of Dino [5] and HuBERT [19] backbones, showing good results on semantic segmentation and retrieval tasks. Similar to DenseAV, we pursue fine-grained modal alignment but with a simpler approach that extends CAV-MAE with dedicated components for contrastive and reconstruction tasks.

3. CAV-MAE Sync

3.1. Overview

Our method employs the contrastive masked autoencoder framework [14], 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.

3.2. Background: CAV-MAE

Given a video consisting of a set of frames and a respective audio signal, CAV-MAE uniformly samples frames from each video and selects for each training step one frame-audio pair consisting of a random frame v_i and the visual representation of the full Mel spectrogram of the respective audio signal a_i. The respective two 2D inputs are then patchified. Then a portion of patches is randomly masked in each modality and a convolutional projection layer tokenizes the remaining frame and audio patches into sequences of visual and audio tokens respectively, also adding a modality type and a positional embedding.

The sequences of unmasked visual tokens u_v and audio tokens u_a are forwarded through separated encoders E_v and E_a to learn modality-specific representations z^v and z_a. Note that, while both encoders share the same ViT architecture and are initialized from identical pretrained weights, they are trained independently without weight sharing.

After the modalities are encoded individually, their interactions are captured in a joint layer J, which is trained through three separate forward passes with shared transformer weights but individual layer normalizations, first one pass for the visual representations alone, second for the audio representations alone. The output patches of those two separate forward passes h^a and h^v are then averaged to form global representations of audio and visual modalities, c_j^a and c_i^v, and used to compute the contrastive loss between the two modalities. The contrastive loss is then defined as:

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

with $s_{i,j} = \|c_i^v\|^T\|c_j^a\|$ being the cosine similarity between normalized visual and audio representations, and $\tau > 0$ a temperature parameter of the similarity distribution.

Finally, for the third pass, the visual and audio tokens are concatenated into a single sequence. This joint representation is used for the masked autoencoding objective. The reconstructed patches are denoted as $\hat{m}_i^a$ for audio and $\hat{m}_i^v$ for video patches for each masked position i. The reconstruction terms L_i^a and L_i^v compute the mean squared error between predicted and original masked patches for audio and visual modalities, respectively:

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

$$L_i^v = \frac{\sum_{i\in|m_v|}\left(\hat{m}_{v_i} - m_{v_i}\right)^2}{|m_v|} \tag{3}$$

The reconstruction loss averages these terms over the batch:

$$L_r = \frac{1}{N}\sum_{i=1}^{N}\left(L_i^a + L_i^v\right) \tag{4}$$

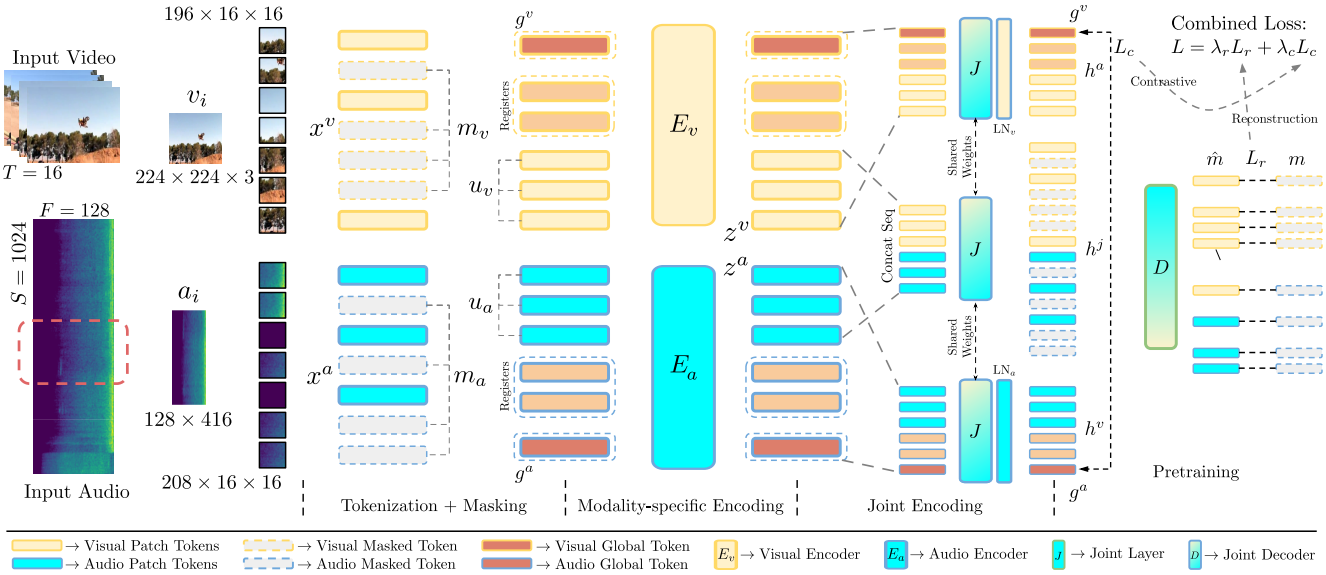

Figure 2. Overview of our approach. Our model processes video frames and audio segments in parallel through separate encoders E_a and E_v, with the audio encoder E_a operating on finer temporal granularity to better align with visual frames. Both modalities interact through the Joint Layer L and the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

The final learning objective balances the contrastive and the reconstruction loss using weighting parameters λ_r and λ_c as $L = \lambda_c L_c + \lambda_r L_r$. This weighted objective ensures that the model learns both modality-specific features through reconstruction and aligned cross-modal representations through contrastive learning.

3.3. Improving Temporal Granularity

We argue that the current contrastive matching of a full audio sequence to a single random frame is a rather loos contrastive objective, as 1) frames from different scenes will be mapped to the same audio as long as they come from the same video and 2) as longer audio usually also contain more than one audio class (e.g. in case of AudioSet) it not only maps several frames to the same audio but also to an audio encoding containing different classes. We therefore first aim to increase the temporal granularity to achieve a more precise audio-visual alignment. To this end, we extract audio segments corresponding to individual video frames. Unlike using a single audio spectrogram for the entire video, this method leverages the natural temporal alignment between audio and visual information and ensures that each audio segment is temporally aligned with its respective video frame, enhancing coherence between modalities.

Temporal Alignment Process. Given a video with T frames and its corresponding audio spectrogram of length S, we extract a fixed-length spectrogram segment of size s_{length} for each frame i. Since video frames and audio spectrogram samples are extracted at different rates from the same time interval, we map each frame index to its corresponding position in the spectrogram using $s_{\text{center}_i} = \lfloor i \cdot S/T \rfloor$. We then extract a window centered at this position, adjusting the boundaries to handle edge cases. The segment extracted from the spectrogram is indexed from a starting position s_{start_i} to an ending position s_{end_i}, where $s_{\text{start}_i} = s_{\text{center}_i} - \lfloor s_{\text{length}}/2 \rfloor$ and $s_{\text{end}_i} = s_{\text{start}_i} + s_{\text{length}}$.

3.4. Disentangling Joint Modality Encoding

In the original architecture [14], patches are optimized for both contrastive and autoencoder objectives using a shallow joint layer, which can hinder the model's ability to learn distinct representations for each objective. To address this, we propose strategies to disentangle these objectives, enhancing the model's performance and representation quality.

Global Token Integration. While traditional MAE approaches aggregate patch tokens to form global representations for downstream tasks [12, 14], we instead introduce dedicated global tokens for the contrastive objective. By separating the global representation pathway from the patch tokens, we reduce the information burden on patch tokens, which can now focus on reconstruction while the global tokens aggregate information during both single-modality and joint encoding stages. We denote these global tokens as g^a and g^v for audio and visual modalities respectively. These tokens serve as global representations of their respective modalities and are used for the contrastive objective and downstream tasks. While these tokens are optimized primarily through the contrastive loss, the entire model's weights, including the encoders and joint layers, are still updated through backpropagation from both objectives.

Register Tokens. Vision transformers often contain high-norm patch tokens that act as computation nodes rather than visual features in the self-supervised setting [10]. We incorporate the idea of register tokens, which helps our method in two ways: It maintains the patch tokens dedicated to the reconstruction objective and allows the global tokens to focus on the contrastive objective. This disentanglement improves the model's ability to capture semantic information and perform localization. These register tokens are appended to u_v and u_a and are processed through the joint layer in the same manner as the global tokens.

Adaptation of the Joint Layer. With the addition of global and register tokens, the joint layer is refined to handle modality-specific representations more effectively. The contrastive loss L_c now exclusively utilizes the global tokens (g^a and g^v), computing similarity scores as $s_{i,j} = \|g_i^v\|^T \|g_j^a\|$. This ensures that the contrastive objective operates on high-level modality representations, while the patch tokens remain to address the reconstruction objective. By disentangling the contrastive and autoencoder objectives, the model can better optimize each task independently. The global tokens provide robust representations for cross-modal alignment, while the patch tokens focus on accurate reconstruction. This separation leads to improved performance in both representation learning and downstream tasks, as the model leverages specialized features for each objective without mutual interference.

3.5. Downstream Tasks

3.5.1. Cross-Modal Retrieval

For cross-modal retrieval, the goal is to retrieve relevant audio segments given a visual query and vice versa. Unlike approaches that use global tokens per modality, we leverage multiple temporal tokens to capture fine-grained relationships between audio and visual data. For each video, we forward all frames and their corresponding audio segments through their respective encoders and joint layer. Using the definitions from Section 3.4, we obtain the final global tokens g^v and g^a after passing through the joint layer J with their corresponding layer normalizations LN_v and LN_a.

Similarity Calculation. For video-to-audio retrieval, consider a query video with a set of visual global tokens $V_q = \{g_1^v, ..., g_T^v\}$ and a target video with audio global tokens $A_t = \{g_1^a, ..., g_T^a\}$, with T as number of temporal tokens per modality. We construct a similarity matrix $S = V_q A_t^\top$ between the two sets where each element $s_{i,j}$ represents the similarity between the i-th visual token of the query video and the j-th audio token of the target video. We compute the final similarity score by averaging the diagonal of S, emphasizing the modality temporal alignment:

$$\text{Similarity Score} = \frac{1}{T} \sum_{t=1}^{T} s_{t,t} \tag{5}$$

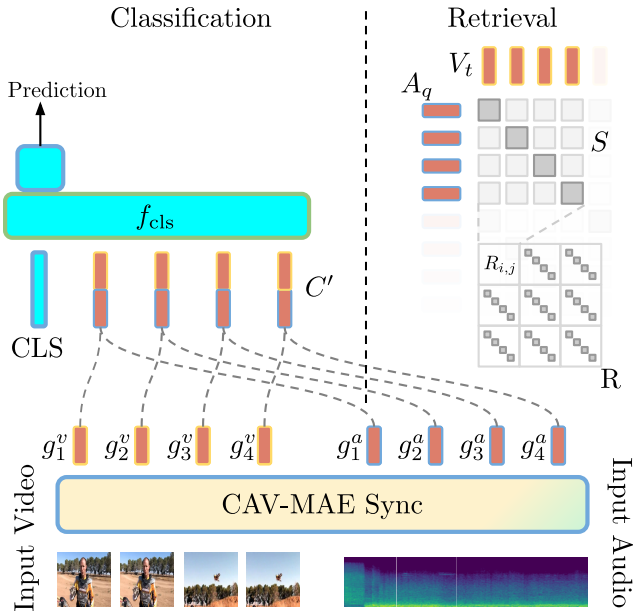

Figure 3. Illustration of our downstream tasks: (1) Classification: using CLS token with f_{cls} projection for video-level prediction, and (2) Retrieval: computing similarity matrix R between audio query A_q and video candidates V_t for cross-modal matching.

This diagonal-focused approach ensures that temporally corresponding token pairs contribute most strongly to the similarity score, promoting retrieval based on both semantic and temporal alignment. For a batch of videos, we compute the similarity scores between all query-target pairs to construct a ranking matrix R, where each element $R_{i,j}$ represents the similarity score between query video i and target video j. The rankings are determined by sorting the similarity scores for each query, with higher scores indicating better matches. Figure 3 illustrates both our classification and cross-modal retrieval tasks.

3.5.2. Classification

For the classification task, we extend the sampling strategy to ensure comprehensive temporal coverage of each video. Instead of loading a single audio-visual segment per video instance, we sample all frames and their corresponding audio segments, effectively increasing the batch size to $B \cdot T$.

For classification, we first obtain the global tokens g^v and g^a from the visual and audio encoders for each temporal step t in video i. These tokens are concatenated at each timestep to form a sequence C_i of length T containing the unified audio-visual representations. We prepend a learnable CLS token to the sequence C_i to aggregate temporal information, resulting in the final sequence $C_i' = \{\text{CLS}, C_{i,1}, \ldots, C_{i,T}\}$ that is passed to the classification head. Let f_{cls} denote our classification head - a two-layer transformer followed by a linear projection. Given the ex-

		AudioSet Eval Subset			VGGSound Eval Subset				AudioSet Eval Subset			VGGSound Eval Subset		
Visual → Audio	**Baselines**	R@1	R@5	R@10	R@1	R@5	R@10	Audio → Visual	R@1	R@5	R@10	R@1	R@5	R@10
	VAB-Encodec [37]	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 [14]	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$^{\text{Scale+}}$ [14]	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 [44]	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 [24]	19.7	–	–	19.0	–	–		17.6	–	–	20.4	–	–
	ImageBind [13]	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.

tended sequence C_i', the classification head produces predictions $\hat{y}_i = f_{\text{cls}}(C_i')$. For multi-class tasks like AudioSet, we use binary cross-entropy loss. For single-class tasks like VGGSound, we use standard cross-entropy loss. In both cases, the model learns to aggregate temporal information over the concatenated audio-visual tokens.

3.5.3. Sound-Prompted Semantic Segmentation

For sound-prompted semantic segmentation, our goal is to make our model output a localization map in a frame given an audio query. We extract the global audio token g^a and all visual tokens h^v and compute the cosine similarity between each h^v and g^a, forming the similarity matrix L corresponding to the 14×14 patches grid of the frame. We then upscale this matrix to the original frame resolution of 224×224 pixels and use it as our predicted localization map.

4. Evaluation

4.1. Datasets

AudioSet. The full AudioSet-2M dataset contains over 2 million 10-second YouTube video clips annotated with 527 audio event classes, which we use for pre-training. For downstream evaluation, we use AudioSet-20K [11], a balanced subset containing $20,000$ samples. For retrieval, we use the subsampled split provided by [14].

VGGSound. The dataset [6] consists of $200,000$ 10-second YouTube videos annotated with 309 classes. Each video contains a visually evident sound source, verified through a pretrained vision classifier. This property makes VGGSound less noisy in terms of audio-visual correspondence.

ADE20K_Sound. This dataset contains 106 images from ADE20K [43] paired with corresponding sound clips from VGGSound. The images and sounds span 20 ADE20K classes, with each image containing objects that produce the paired sound (e.g., an image of a dog paired with a barking sound). The dataset was created by manually selecting ADE20K images and matching them with semantically relevant audio clips from VGGSound.

4.2. Downstream Tasks

To assess the capabilities of our proposed approach, we evaluate performance on three key downstream tasks. Further implementation details can be found in the supplement.

Zero-shot Audio-Visual Retrieval. Given a query from one modality, the model must retrieve the corresponding sample from the other modality. We evaluate bidirectional retrieval (Visual→Audio and Audio→Visual) using Recall@k metrics ($k = \{1, 5, 10\}$) on AudioSet and VGGSound test sets. For fair comparison, we follow the evaluation protocol and subsampling from CAV-MAE [14], using cosine similarity between embeddings to rank candidates.

Sound-Prompted Image Segmentation. Following [16], we evaluate cross-modal localization using ADE20K_Sound. Given an audio prompt, the model must segment corresponding regions in the paired image. Performance is measured via mean Average Precision (mAP) and mean Intersection over Union (mIoU) across 20 classes.

Classification. We assess representation quality through linear probing on AudioSet and VGGSound classification tasks. Following standard practice, we freeze the pretrained encoder and train only a linear classifier, using mean Average Precision (mAP) for AudioSet's multi-label case and accuracy for VGGSound's single-label setting.

4.3. Comparison to State-of-the-Art

Zero-shot Retrieval. We evaluate our model's retrieval performance in both directions - Visual to Audio (V→A) and Audio to Visual (A→V) - on AudioSet and VGGSound datasets, following the same subsampling strategy as [14]. For baselines, we compare against state-of-the-art audio-visual models including CAV-MAE [14], ImageBind [13], AVSiam [24], and VAB [37] as a reference for a model fine-tuned for retrieval serving as an upper bound. We report numbers from original papers and otherwise use officially released checkpoints where available. The retrieval metrics are computed using cosine similarity between query and target embeddings as detailed in Section 3.5. As shown in

Baselines	Pretrain Dataset	AS20K↑	VGGSound↑
VAB-Encodec [37]	AS-2M + VGGS	33.3	57.6
CAV-MAE [14]	AS-2M	27.3	-
CAV-MAE$^{\text{Scale+}}$ [14]	AS-2M	28.5	47.7
CAV-MAE$^{\text{Scale++}}$ [14]	AS-2M	29.2	51.1
CAV-MAE$^{\text{Scale+++}}$ [14]	AS-2M	25.3	51.6
MaViL [20]	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.

Table 1, our model achieves state-of-the-art performance in both directions, suggesting that our model learns a balanced joint embedding space. These results demonstrate that enforcing temporally consistent audio-visual correspondences during pretraining together with a disentangling of the contrastive MAE objective enables better generalization to downstream retrieval tasks.

Classification. While our model achieves strong retrieval performance through finer-grained temporal alignment, we also evaluate its representation learning capabilities through linear probing. Prior work has observed trade-offs between retrieval and representation learning performance, where optimizing for one task often comes at the expense of the other. Our goal is to maintain strong performance across all tasks through a unified architecture rather than specializing in a single objective. Table 2 compares linear probing performance on AudioSet-20K and VGGSound classification. Our model achieves 30.5 mAP on AudioSet and 52.7% accuracy on VGGSound, outperforming CAV-MAE variants and MaViL when using only AudioSet-2M pretraining.

Sound-Prompted Image Segmentation. Following [16], we finally assess our model's ability to localize sound sources in images using the ADE20K_Sound dataset. This dataset pairs 106 images from ADE20K with corresponding sound clips from VGGSound, spanning 20 ADE20K classes. The task requires the model to segment regions in an image corresponding to a given sound prompt, effectively testing cross-modal capabilities. Performance is measured using mean Average Precision (mAP) and mean Intersection over Union (mIoU), averaged across the 20 ADE20K classes. As shown in Table 3, our model achieves 22.7 mIoU on sound-prompted segmentation, performing on par with previous self-supervised approaches like CAV-MAE and ImageBind. Note that while the current best model, DenseAV, achieves higher performance (24.2 with DinoV2+LoRA), we only directly compare to approaches with identical or similar backbones and architecture.

4.4. Ablation Studies

We conduct a series of ablation studies to evaluate the impact of different components on our model's performance. Tables 4–8 summarize the results of these experiments.

Baselines	mAP ↑	mIoU ↑
DAVENet [17]	16.8	17.0
CAVMAE [14]	26.0† / 21.2	20.5† / 20.9
ImageBind [13]	18.3	19.1
Ours	**22.6**	**22.7**

Table 3. Sound-prompted semantic segmentation: Comparison of sound localization methods on ADE20K Sound dataset [16]. †Original reported by [16] / our reproduction.

Visual → Audio	AudioSet Eval Subset			VGGSound Eval Subset		
	R@1	R@5	R@10	R@1	R@5	R@10
CAV-MAE$^{\text{Scale+}}$	15.7	35.2	45.3	11.1	26.5	35.0
↪ *Increase batch size 128 → 256*						
CAV-MAE$^{\text{Scale++}}$	19.7	40.2	50.7	14.8	31.6	41.0
↪ *Pretrain and retrieve with 16 temporal tokens using diagonal similarity (Sec. 3.5)*						
CAV-MAE$^{\text{Scale++}*}$	23.9	46.8	58.0	16.1	36.2	46.1
↪ *Increase batch size 256 → 512 and λ_c=0.1*						
CAV-MAE$^{\text{Scale+++}}$ (λ_c=0.1)	30.1	54.9	64.5	21.0	44.0	56.8
Ours	**35.2**	**58.3**	**67.6**	**27.9**	**51.7**	**61.8**

Table 4. Retrieval performance comparison showing the progression from CAV-MAE to our approach. The results demonstrate that simply using temporal tokens with diagonal similarity yields weaker performance than ours, highlighting the importance of our global and register tokens combined with fine-grained pretraining.

Establishing a Strong CAV-MAE Baseline. Table 4 shows our systematic optimization of CAV-MAE as a baseline. Starting with CAV-MAE$^{\text{Scale+}}$, increasing batch size to 256 yields CAV-MAE$^{\text{Scale++}}$ with improved AudioSet retrieval (R@1 from 18.8 to 19.7). Using 16 temporal tokens during both pretraining and retrieval, along with diagonal sequence similarity for retrieval (CAV-MAE$^{\text{Scale++}*}$), further boosts performance to 23.9 R@1. Finally, larger batches (512) and stronger contrastive weight ($\lambda_c = 0.1$) in CAV-MAE$^{\text{Scale+++}}$ achieves 30.1 R@1. For fair comparison, our model uses the same hyperparameters with 4s audio segments, reaching 35.2 R@1 on AudioSet and 27.9 R@1 on VGGSound with the same parameter count and a fraction of the token count.

# Frames		AudioSet Eval Subset			VGGSound Eval Subset		
Train	Eval	R@1	R@5	R@10	R@1	R@5	R@10
10	10	31.2	55.3	65.6	25.6	49.6	59.3
10	16	34	57.3	66.8	27.0	50.7	60.6
16	10	32.3	55.0	65.3	26.3	48.7	59.6
16	16	**35.2**	**58.3**	**67.6**	**27.9**	**51.7**	**61.8**

Table 5. Effect of frame sampling density on retrieval performance during pre-training and evaluation. Results demonstrate consistent improvements with denser temporal sampling across both stages.

Audio Length Ablation. To promote fine-grained temporal alignment between audio and visual modalities, we investigate the optimal temporal granularity for audio sam-

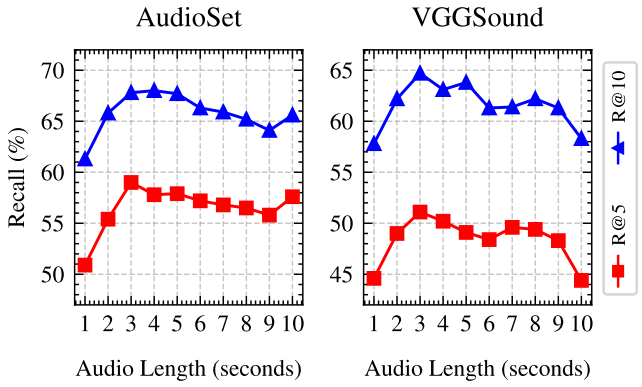

Figure 4. Impact of audio segment length on model performance. Experiments show 3-4 second segments achieve optimal results while reducing computational costs compared to standard 10-second segments.

# Regs.	AudioSet Eval Subset			VGGSound Eval Subset			Localization	Classification
	R@1	R@5	R@10	R@1	R@5	R@10	mIoU	AS20k (mAP)
0	**35.5**	59.0	68.0	27.9	50.9	62.4	20.9	27.1
4	32.9	58.1	67.0	**28.2**	51.3	63.2	21.7	27.5
8	35.2	58.3	67.6	27.9	51.7	61.8	22.8	**30.8**
16	34.9	**59.2**	**68.5**	27.7	**52.2**	**64.0**	**22.9**	26.8

Table 6. Ablation studies on the number of registers. Results demonstrate consistent improvements with denser temporal sampling across both stages.

Global	AudioSet Eval Subset			VGGSound Eval Subset			Localization	Classification
	R@1	R@5	R@10	R@1	R@5	R@10	mIoU	AS20k (mAP)
w/o	30.1	55.3	64.9	21.6	48.1	60.7	20.7	26.6
w/	**35.2**	**58.3**	**67.6**	**27.9**	**51.7**	**61.8**	**22.8**	**30.8**

Table 7. Effect of global token on model performance. The global token significantly improves cross-modal retrieval on both datasets while enhancing localization and classification tasks, demonstrating its importance for capturing audio-visual relationships.

Ratio	AudioSet Eval Subset			VGGSound Eval Subset			Localization
	R@1	R@5	R@10	R@1	R@5	R@10	mIoU
0.6	**39.1**	**63.3**	**71.7**	**28.6**	**55.5**	**68.7**	19.1
0.75	35.2	58.3	67.6	27.9	51.7	61.8	22.8
0.9	21.5	42.3	53.5	16.7	37.9	50.1	**23.4**
multi {*0.6-0.9*}	27.3	51.6	62.1	22.1	45.7	59.9	21.2

Table 8. Effect of masking ratio on retrieval and localization performance. Trade-off analysis between different masking ratios shows 0.75 provides optimal balance between tasks.

pling, as videos typically contain multiple distinct audio events that should be precisely matched with their corresponding visual frames. In Figure 4, our experiments show that 3-4 second segments provide optimal performance for our architecture and pre-training approach, outperforming the standard 10-second segments in zero-shot retrieval tasks. Note that this shorter duration also brings computational benefits - since our model samples one random frame and audio segment during pre-training, we process only 30-40% of the audio tokens compared to using full 10-second segments. Audio segments shorter than 3 seconds lead to degraded performance, suggesting this could be a limit for effective learning in our framework.

Number of Frames. Table 5 investigates the impact of temporal sampling density during both pre-training and evaluation. While increasing frames from 10 to 16 in either stage improves performance, the best results come from denser sampling in both, with pre-training and evaluating with 16 frames yielding the best score.

Number of Registers. In Table 6, we assess the impact of varying register counts. While registers are known to boost classification performance in vision transformers [10], optimal counts can vary by task. We observe that increasing the number of registers consistently improves localization performance, from 20.9 IoU with zero registers to 22.9 with 16 registers. We chose 8 register tokens as our standard setting since it achieved competitive performance across all tasks, supporting our goal of a balanced model.

Global Token. The presence of a global token is evaluated in Table 7. Incorporating the global token enhances the performance for retrieval on AudioSet (+3.6% on average) and VGGSound (+3.7% on average), while also improving localization (mIoU from 20.7 to 22.8) and classification (AS20k mAP from 26.6 to 30.8). These consistent improvements across all tasks highlight the global token's role in capturing global context and helping disentangle the contrastive and reconstruction objectives.

Masking Ratio. Table 8 examines different masking ratios during pretraining. A lower masking ratio of 0.6 achieves the highest retrieval performance but results in poorer localization (IoU 19.1). A higher ratio of 0.9 significantly degrades retrieval performance but improves localization (IoU 23.3). A ratio of 0.75 provides a good balance, with strong retrieval and localization (IoU 21.8) performance. Applying a multi-ratio strategy [24] led to worse performance across all metrics. Therefore, we use 0.75 masking in our final model, as it provides the best trade-off between retrieval and localization capabilities.

5. 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.

Acknowledgements

Edson Araujo is supported by German Federal Ministry of Education and Research (BMBF) project STCL - 01IS22067. This research is in part supported by the MIT-IBM Watson AI Lab.

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. 1, 2

[3] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *European Conference on Computer Vision*, pages 435–451, 2018. 2, 3

[4] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in Neural Information Processing Systems*, 29, 2016. 1, 2

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725, 2020. 2, 6, 1

[7] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 3

[8] 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

[9] 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

[10] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 5, 8

[11] 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. 2, 6, 1

[12] 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. 2, 4

[13] 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. 3, 6, 7

[14] 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, 6, 7

[15] 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, 2

[16] 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. 3, 6, 7

[17] 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. 3, 7

[18] 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

[19] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 3

[20] 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, 2, 7

[21] Jongsuk Kim, Hyeongkeun Lee, Kyeongha Rho, Junmo Kim, and Joon Son Chung. EquiAV: Leveraging Equivariance for Audio-Visual Contrastive Learning. In *ICML*, pages 24327–24341, 2024. 3

[22] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[23] 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

[24] Yan-Bo Lin and Gedas Bertasius. Siamese vision transformers are scalable audio-visual learners. In *ECCV*, 2024. 1, 2, 6, 8

[25] 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

[26] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *International Conference on Learning Representations*, 2020. 2

[27] 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

[28] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12945, 2021. 2

[29] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12475–12486, 2021. 2

[30] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision*, pages 631–648, 2018. 2

[31] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision*, pages 801–816, 2016. 2

[32] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *IEEE/CVF International Conference on Computer Vision*, pages 9577–9587, 2021. 3

[33] Adria Recasens, Pauline Luc, Jean-Baptiste Alayrac, Luyu Wang, Florian Strub, Corentin Tallec, Mateusz Malinowski, Viorica Pătrăucean, Florent Altché, Michal Valko, et al. Broaden your views for self-supervised video learning. In *IEEE/CVF International Conference on Computer Vision*, pages 1255–1265, 2021. 3

[34] 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, 2

[35] Pritam Sarkar and Ali Etemad. Self-supervised audio-visual representation learning with relaxed cross-modal synchronicity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9723–9732, 2023. 3

[36] 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

[37] 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. 2, 6, 7

[38] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2023. 1, 2

[39] Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, and Aaron van den Oord. Multimodal self-supervised learning of general audio representations. *arXiv preprint arXiv:2104.12807*, 2021. 3

[40] 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

[41] Zhaoyang Zeng, Daniel McDuff, Yale Song, et al. Contrastive learning of global and local video representations. *Advances in Neural Information Processing Systems*, 34: 7025–7040, 2021. 3

[42] 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

[43] 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. 2, 6

[44] 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. 3, 6