Siamese Vision Transformers are Scalable Audio-visual Learners

Yan-Bo Lin and Gedas Bertasius

Department of Computer Science
University of North Carolina at Chapel Hill
{yblin,gedas}@cs.unc.edu

Abstract. Traditional audio-visual methods rely on independent audio and visual backbones, which is costly and not scalable. In this work, we investigate using an audio-visual siamese network (AVSiam) for efficient and scalable audio-visual pretraining. Our framework uses a single shared vision transformer backbone to process audio and visual inputs, improving its parameter efficiency, reducing the GPU memory footprint, and allowing us to scale our method to larger datasets and model sizes. We pretrain our model using a contrastive audio-visual matching objective with a multi-ratio random masking scheme, which enables our model to process larger audio-visual instance batches, helpful for contrastive learning. Unlike prior audio-visual methods, our method can robustly handle audio, visual, and audio-visual inputs with a single shared ViT backbone. Furthermore, despite using the shared backbone for both modalities, AVSiam achieves competitive or even better results than prior methods on AudioSet and VGGSound for audio-visual classification and retrieval. Our code is available at https://github.com/GenjiB/AVSiam

1 Introduction

The last few years have witnessed remarkable progress in audio-visual representation learning [32, 34, 64, 65, 72, 74, 80]. However, most modern audio-visual approaches use separate audio and visual backbones and costly audiovisual pretraining protocols, limiting their scalability. For example, the recent state-of-the-art audio-visual method MAViL [28] requires 5,120 V100 GPU hours for pretraining, which is not feasible for many research labs. Also, the best-performing variant of audio-visual MBT [52] uses more than 48GB of GPU memory during training, which requires costly A100 or H100 GPU servers, unavailable to many researchers. These factors make it difficult to scale these audio-visual models to larger datasets and bigger models, which has been the recent trend in many computer vision and multimodal modeling domains [12, 17, 69, 79].

In addition to large pretraining and GPU memory costs, the modality-specific backbones used by conventional audio-visual methods have several other shortcomings. First, designing separate audio and visual backbones may lead to handcrafted priors and inductive biases, which not only contribute to increased research/engineering effort of optimizing models for a specific modality but can

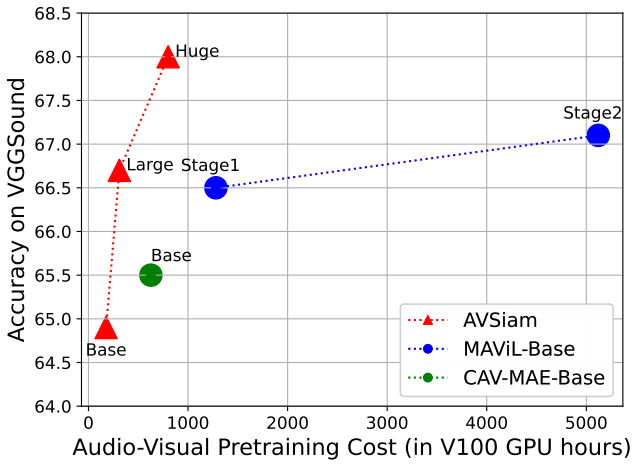

Fig. 1: Our Audio-visual Siamese network (AVSiam) uses a single shared backbone to process audio and visual data, which reduces its GPU memory footprint and allows us to scale our method to larger datasets and model sizes. Compared to prior audio-visual approaches [26, 28], which are very costly, our model is both more efficient and also achieves higher accuracy on standard audio-visual classification benchmarks.

also be detrimental to data-driven representation learning [63, 68, 81]. Second, approaches relying on modality-specific backbones lack the flexibility to handle the cases of variable inputs and missing modalities (e.g., visual-only, audio-only, audio-visual, etc.). Lastly, the modality-specific audio-visual models are not parameter efficient, which increases their memory footprint and limits their scalability to larger datasets and bigger models.

Motivated by these observations, we introduce an Audio-visual Siamese network (AVSiam) for efficient and scalable audio-visual pretraining. Our method is inspired by several recent works showing the ability of Vision Transformers (ViTs) to generalize to different modalities and domains, including audio and video [20, 24, 26, 28, 39, 62]. Also, since audio can be represented as a 2D spectrogram with a spatial 2D structure akin to *audio images*, modern vision architectures (e.g., ViTs or CNNs) have been shown to process such audio images effectively. Despite these findings, most modern audio-visual methods [14, 20, 26, 28, 38, 39, 46] rely on modality-specific backbones for audio and visual data processing, since individually tailored models typically lead to better results on various audio-visual benchmarks.

Instead, we investigate using a single Siamese vision transformer to process both audio and visual inputs, leading to a unified audio-visual model with a reduced GPU memory footprint. To pretrain our model, we use a contrastive audio-visual matching objective with a novel multi-ratio random masking scheme, which randomly masks audio and visual patches at different ratios. Unlike fixed ratio masking [35], our multi-ratio random masking enables the model to learn

robust representations across a spectrum of available information. Such a masking scheme and a shared audio-visual backbone also allow us to consider larger audiovisual instance batches, benefiting the contrastive learning process.

Our experimental results show that AVSiam can robustly process audio-only, visual-only, or audio-visual inputs with a single shared backbone. Furthermore, as depicted in Figure 1, despite using a shared backbone for both modalities, AVSiam achieves competitive or superior results than prior methods with separate audio and visual backbones on audio-visual classification and retrieval benchmarks, including AudioSet-20K [19], AudioSet-2M [19], VGGSound [11], and MSR-VTT [76] while requiring significantly fewer resources for pretraining (i.e., **28.9**× faster than MAViL) Lastly, the efficiency of our approach enables training our model on larger datasets or with backbones of increased capacity, leading to further improvements.

2 Related Work

2.1 Audio-Visual Representation Learning

Audio-visual data provides a rich supervisory signal for learning audio and visual representations for the tasks of audio-visual event localization [45,56–58,65, 73,77], audio-visual video parsing [14,40,48,64,72], and audio-visual classification [20,25,26,28,47,52,69]. With the growing availability of audio-visual data on the Web, there has been a lot of progress in self-supervised audio-visual representation learning [43,49,54]. The methods in [2–7,43,49,50,55] exploit the natural synchronization between audio and visual data by learning to predict whether a given audio and video pair matches. Additionally, audio-visual contrastive learning methods [13,31,43,44] learn to associate audio-visual features from the same videos and differentiate representations from all other instances in a mini-batch. Recently, masked autoencoders (MAE) [8,9,20,26,28,42,60,62,75] have demonstrated the capability not only to learn robust visual features but also audio representations. Audio-visual MAE approaches [20,26,28] reconstruct the original audio and visual data from masked audio-visual tokens to learn correlations between audio and images/video. However, while effective, such MAE-based approaches [20, 26, 28] typically require a high computational cost, which limits their scalability and adaptation. To address these issues, we propose AVSiam, an audio-visual model that uses a single backbone for audio and visual data to reduce the cost of audio-visual modeling while still attaining solid performance.

2.2 Unified Multimodal Representation Learning

The versatile design of a Transformer [16] has made it easy to process multimodal data (e.g., sound, images, videos, text, etc.), thus unlocking the potential for diverse multimodal applications [1,21,22,37,38,41,59]. Several recent attention-based architectures [10,29,30,66] have enabled direct processing of data across different modalities. Another strategy for multimodal data processing involves augmenting models with modality-specific weights attached to a single

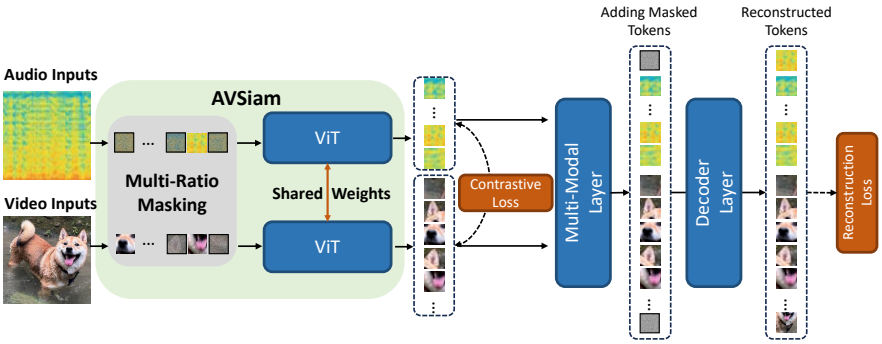

Fig. 2: Our Pretraining Framework. Our AVSiam approach uses a single shared vision transformer backbone to process both audio and visual data. To train our model, we use a novel multi-ratio masking scheme, which randomly masks audio and visual tokens at various masking ratios. As our pretraining objectives, we employ audio-visual contrastive matching and audio-visual token reconstruction loss functions.

modality-agnostic backbone [1, 15, 37, 61]. Furthermore, the models with shared weights across different tasks [46, 51, 78] recently achieved impressive results while demonstrating the flexibility of such approaches. However, most existing audio-visual approaches [20, 26, 28, 52] still rely on modality-specific model design, which makes them costly and limits their scalability. Instead, in this work, we introduce AVSiam, a framework that uses a single audio-visual backbone for learning effective audio-visual representations in an efficient manner.

3 Technical Approach

In this section, we present AVSiam, an Audio-visual Siamese network that uses a shared-weight vision transformer (ViT) backbone to process audio and visual inputs for efficient audio-visual pretraining. We illustrate our pretraining framework in Figure 2 and describe it in more detail below.

3.1 The AVSiam Model

Audio-Visual Input Embeddings. For the visual inputs, we use an RGB video frame $I \in \mathbb{R}^{H_v \times W_v \times 3}$, representing a frame randomly selected from the video at time t with spatial dimensions $H_v \times W_v$. For the audio inputs, we process an audio spectrogram $A \in \mathbb{R}^{H_a \times W_a}$, covering approximately 10 seconds. Following ViT [16], we patchify a RGB frame I into n non-overlapping patches, and transform them into visual embeddings $\mathbf{X}_v \in \mathbb{R}^{n \times d}$ Similarly, an audio spectrogram A is projected into audio embeddings $\mathbf{X}_a \in \mathbb{R}^{k \times d}$ with k patches. Note that due to different audio and visual channel dimensions, to obtain audio embeddings, we average the weights of the 3-channel projection layers of the pretrained ViT into single-channel weights when processing audio inputs.

Model Architecture. We use a standard ViT [16] architecture to implement our AVSiam model. The main difference compared to prior transformer-based audio-visual approaches [26, 28, 52] is that our model shares the full set of parameters of a pre-trained ViT across both audio and visual streams. We use a shared audio-visual encoder to process audio and visual inputs $\mathbf{X}_a$ and $\mathbf{X}_v$ and then pool the outputs for each modality via average pooling, which produces audio and visual features $\mathbf{F}_a \in \mathbb{R}^d$ and $\mathbf{F}_v \in \mathbb{R}^d$, respectively. As for the multimodal layers and decoder, we follow the implementation of CAV-MAE [26] and MAE [27]. Specifically, the audio $\mathbf{F}_a$ and visual $\mathbf{F}_v$ tokens are concatenated and fed into a multimodal layer MM(.), which processes audio and visual tokens via a joint two-layer self-attention block. During training, we also use a six-layer self-attention decoder Dec(.) to reconstruct the masked audio and visual tokens as shown in Figure 2. We note that even though our AVSiam is based on the ViT design, in practice, it can be easily adapted to other backbones.

3.2 Training the AVSiam Model

Multi-Ratio Input Masking. Prior work [35, 36, 71] has shown that reducing the number of input tokens can lead to dramatic savings in computational resources (i.e., GPU memory usage, pretraining time, etc.). Since audio spectrogram inputs typically have substantially more tokens (patches) than images (i.e., 512 vs. 196), such computational savings can be even more significant in the audio-visual domain. Although masking a large portion of tokens enables efficient large-scale pretraining, deciding a masking ratio that provides optimal efficiency vs. accuracy trade-off is not trivial. For example, a high masking ratio can save GPU memory, leading to bigger batch sizes and, thus, potentially better performance for contrastive learning-based methods. On the other hand, a high masking ratio also leads to a significant loss of information, which could negatively affect the quality of learned representations.

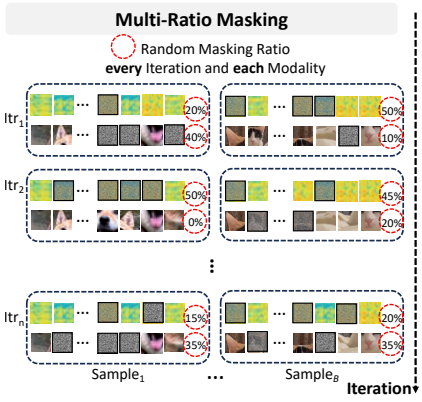

Fig. 3: Multi-Ratio Masking. We apply random masking to audio and visual tokens in various proportions during each training iteration.

To incorporate the benefits of different masking ratios, we randomly mask audio and visual patches at varying ratios during each training iteration. As depicted in Figure 3, during each pretraining iteration, every audio and visual instance in a mini-match will be randomly assigned a masking ratio from 0% to 50%. Unlike the fixed ratio masking approach [35], which always uses the same number of tokens, our multi-ratio random masking scheme learns from a diverse number of tokens, which leads to more robust audio-visual representations. To efficiently implement our proposed masking scheme, we first randomly partition

samples in a batch into groups. All samples belonging to the same group are assigned the same masking ratio from a predefined range of 0% to 50% in 5% increments. The masking ratio is randomly selected for each group. Such a sample partitioning and masking scheme ensures that all samples in the same group have an identical number of unmasked tokens and, thus, can be stacked together for efficient GPU memory utilization.

Audio-visual Pretraining Objectives. We use audio-visual contrastive matching and token reconstruction objectives to pretrain our model. Specifically, we use audio $\mathbf{F}_a$ and visual features $\mathbf{F}_v$, obtained from our shared AVSiam encoder for contrastive audio-visual matching as in [26, 53]:

$$\mathcal{L}_c(\mathbf{F}_a, \mathbf{F}_v) = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{exp(\mathrm{g}(\mathbf{F}_a^i, \mathbf{F}_v^i)/\tau)}{\sum_{j=1}^{B} exp(\mathrm{g}(\mathbf{F}_a^i, \mathbf{F}_v^j)/\tau))}. \tag{1}$$

Here, g(.) is a standard cosine similarity function, and B and τ denote mini-batch size and temperature, respectively. Also, inspired by the recent audio-visual MAE approaches [20, 26, 28], we adopt a masked token reconstruction [27] objective to learn a more effective audio-visual representation. Specifically, the audio-visual decoder Dec(.) processes the masked and unmasked audio and visual tokens as inputs to predict the original audio spectrogram and image as:

$$\tilde{I}, \tilde{A} = \mathrm{D}ec(\mathrm{MM}(\mathbf{F}_a^u, \mathbf{F}_v^u), \mathbf{F}_a^m, \mathbf{F}_v^m), \tag{2}$$

where $\mathbf{F}_a^u, \mathbf{F}_v^u, \mathbf{F}_a^m$, and $\mathbf{F}_v^m$ are unmasked audio, unmasked visual, masked audio, and masked visual tokens respectively. Note that unmasked tokens are extracted without applying average pooling. The reconstructed audio spectrogram $\tilde{A} \in \mathbb{R}^{H_a \times W_a}$ and image $\tilde{I} \in \mathbb{R}^{H_v \times W_v \times 3}$ has the same shape as the original spectrogram and image inputs. The reconstruction objective is written as:

$$\mathcal{L}_{rec} = \frac{1}{B}\sum_{i=1}^{B} (\tilde{A}^i - A^i)^2 + (\tilde{I}^i - I^i)^2. \tag{3}$$

The final objective is obtained by combining the two losses: $\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_c$.

Supervised Finetuning with Mixed Modality Inputs. After pretraining our framework as described above, we proceed with supervised fine-tuning of our model on various downstream audio-visual understanding tasks. Unlike in the pretraining stage, in the supervised fine-tuning stage, we do not mask any of the input tokens and use a standard training protocol as in CAV-MAE [26] and MAViL [28]. To robustly handle missing modalities and ensure generalization to audio-only and visual-only settings, during each iteration of finetuning, we randomly select one of three types of inputs, audio-only, visual-only, or audio-visual, and feed it into our shared ViT encoder. Once our shared encoder processes these inputs, we average all tokens encoded by the ViT to obtain representations for each modality (i.e., $\mathbf{F}_a$ for audio-only, $\mathbf{F}_v$ for visual-only, $\mathbf{F}_a$ and $\mathbf{F}_v$ for audio-visual inputs). The resulting features are then fed into an MLP layer for the final prediction. We use binary cross-entropy loss for AudioSet [19] and cross-entropy loss for VGGSound [11] as in CAV-MAE [26] for our supervised finetuning.

3.3 Implementation Details

We used a standard Vision Transformer (ViT) architecture for the base, large, and huge variants. All encoder layers, adapted from ViT, were pretrained on ImageNet-21K. For the multimodal layers, we utilized the weights from the last two layers of the pretrained ViT to initialize the first two layers of the decoder to process all unmasked audio-visual tokens. As for the decoder, we adhered to the default MAE settings for the base, large. Specifically, the decoder processed all masked and unmasked tokens as inputs where the masked tokens were randomly initialized and learnable. The weights of the decoder were randomly intitialized. For contrastive loss, the temperature was set to $\tau = 0.05$. We utilized two independent loss scalers, one for contrastive loss and the other for the MAE loss. During supervised finetuning with mixed modality inputs (see details above), we randomly sampled either audio or visual inputs in half of our training iterations to train our model on audio-only or visual-only inputs. In the remaining iterations, we used both audio and visual inputs. As for optimization, we used Adam optimizer for both the pretraining and fine-tuning stages. We set the learning rate to $1e-4$ for pretraining. During finetuning, we set $1e-4$, $5e-6$, and $5e-5$ for AudioSet-20K, AudioSet-2M, and VGGSound.

4 Experimental Setup

Audio-visual Classification. We pretrain our model on AudioSet-2M [19]. We note that we could only obtain 1.7M videos due to expired YouTube links. After pretraining, the model is finetuned on three audio-visual classification datasets: (1) AudioSet-20K containing 20K samples from the same domain as the pretraining data (i.e., AudioSet-2M), (2) the original AudioSet-2M dataset, and (3) VGGSound [11]. For evaluation metrics, we use standard mean Average Precision (mAP) for AudioSet-20K and AudioSet-2M and top-1 accuracy for VGGSound.

Audio-visual Retrieval. To evaluate whether our learned audio and visual representations generalize to different tasks, we also evaluate AVSiam on video-to-audio and audio-to-video retrieval tasks without any additional finetuning similar to the setting of CAV-MAE [26]. Additionally, we follow the setup of MAViL [28], and also include audio-to-video retrieval results after finetuning the pretrained model. The recall at 1 metric ($R@1$) is then computed using the cosine similarity between the audio and visual representations. We use the original evaluation splits from CAV-MAE, which include 1,725 and 1,545 videos from AudioSet and VGGSound, respectively. We also include audio-visual retrieval results on the test set of MSR-VTT [76] containing 2,990 videos.

5 Results and Analysis

We first discuss our audio-visual classification results on AudioSet and VGGSound. Afterward, we present our analysis for audio-visual retrieval on AudioSet, VGGSound, and MSR-VTT. Lastly, we discuss our ablation studies.

Table 1: Audio-Visual Event Classification: We compare the results of our proposed AVSiam with previous baselines on audio-visual event classification on the AudioSet and VGGSound benchmarks. We de-emphasize MBT* [52] due to using non-standard training-test splits. Note that AVSiam-Base^{+} uses more data from ACAV [33] dataset for pretraining only.

Method	Audio Encoder	Visual Encoder	V100 Hours	#Param	AS-20K (mAP↑)	AS-2M (mAP↑)	VGGSound (Acc.↑)
Audio-Video Models							
G-Blend [70]	-	-	-	-	37.8	41.8	-
Perceiver [30]	-	-	-	-	-	44.2	-
Attn AV [18]	-	-	-	-	-	44.2	-
Zorro-Swin [59]	-	-	-	161M	-	46.5	-
MBT* [52]	AST-B	ViT-B	-	172M	43.9	49.6	64.1
Masked Autoencoder							
AV-MAE [20]	AST-B	ViT-B	2854	179M	-	50.0	64.2
AV-MAE-Large	AST-L	ViT-L	-	626M	-	51.8	65.0
CAV-MAE [26]	AST-B	ViT-B	672	164M	42.0	51.2	65.5
MAViL-Stage1 [28]	AST-B	ViT-B	1280	172M	44.6	51.9	66.5
MAViL-Stage2	AST-B	ViT-B	5120	172M	44.9	53.3	67.1
Shared-Weight Encoder							
AVSiam-Base	ViT-B (shared)		177	**100M**	41.6	50.1	64.9
AVSiam-Base^{+}	ViT-B (shared)		450	**100M**	43.0	51.4	66.7
AVSiam-Large	ViT-L (shared)		310	332M	44.1	52.1	67.1
AVSiam-Huge	ViT-H (shared)		800	672M	**45.0**	**54.1**	**68.0**

5.1 Audio-Visual Classification Results

In Table 1, we present our audio-visual classification results on AudioSet and VGGSound using multiple axes of comparison, including the types of audio and visual encoders, the pretraining time in V100 GPU hours, the number of parameters, and audio-visual classification accuracy. We compare our AVSiam approach (bottom part of the table) with standard audio-visual models [18, 30, 52, 59, 70] (top part of the table) and also with recent MAE-based approaches [20, 26, 28] (middle part of the table). Based on these results, first, we observe that our default AVSiam-Base model performs similarly as AV-MAE [20] and MBT [52] on AudioSet-2M (i.e., **50.1** mAP vs. **50.0** mAP vs. **49.6** mAP) and VGGSound (i.e., **64.9%** vs. **64.2%** vs. **64.1%**). We also observe that AVSiam achieves worse results than the latest state-of-the-art MAE-based approaches such as CAV-MAE [26] and MaViL [28] on AudioSet-2M (i.e., **50.1** mAP vs. **51.2** mAP vs. **53.3** mAP) and VGGSound (i.e., **64.9%** vs. **65.5%** vs. **67.1%**. However, unlike all these other approaches, which use modality-specific audio and visual encoders, AVSiam uses a single, shared audio-visual backbone, which has a smaller number of parameters (i.e., **100M** vs. **164M** vs. **172M**) and requires significantly less time for pretraining (i.e., **177** vs. **672** vs. **5120** V100 GPU hours. In particular, we note that the best performing MAViL variant requires **5,120** V100 GPU hours for pretraining which is **28.9**× more than our model (i.e., **177** V100 GPU hours). These results highlight the efficiency of our approach.

Next, leveraging the efficiency of our model, we show that we can scale AVSiam to larger training datasets and also larger model variants (i.e., ViT-

Table 2: Zero-shot Audio-Visual Retrieval: We evaluate AVSiam for video-to-audio and audio-to-video retrieval on the AudioSet, VGGSound, and MSR-VTT datasets, all in zero-shot settings. The results are reported using the recall at 1 metric (R@1). Compared to prior approaches, AVSiam achieves the best results across all metrics and datasets while using fewer parameters.

	Audio Encoder	Visual Encoder	#Params (M)	AudioSet V→A	AudioSet A→V	VGGSound V→A	VGGSound A→V	MSR-VTT V→A	MSR-VTT A→V
CAV-MAE	AST-B	ViT-B	164	16.1	13.5	14.7	12.1	4.9	8.3
CAV-MAE$^+$	AST-B	ViT-B	164	18.8	15.1	14.8	12.8	7.6	13.3
AVSiam-Base	ViT-B (shared)		**100**	**19.7**	**17.6**	**19.0**	**20.4**	**9.3**	**16.1**

Large, ViT-Huge). Specifically, we pretrain our model on AudioSet-2M [19] augmented with additional samples from VGGSound (200K) [11] and ACAV (2.4M) [33]. We report these results under the AVSiam-Base$^+$ variant. We observe that scaling our model to a larger dataset size consistently improves the performance across all three datasets (i.e., +**1.4%**, +**1.3%**, and +**1.8%**). Moreover, AVSiam-Base$^+$ outperforms the most efficient MAE baseline, CAV-MAE (i.e., AudioSet-20K: **43.0%** vs. **42.0%**, AudioSet-2M: **51.4%** vs. **51.2%**, and VGGSound: **66.7%** vs. **65.5%**) while still being faster to pretrain (i.e., **450** vs. **672** V100 hours) despite using a significantly larger dataset size.

Afterward, we also investigate scaling our model size by considering the AVSiam-Large and AVSiam-Huge variants. Compared to AVSiam-Base, AVSiam-Large and AVSiam-Huge demonstrate consistent and significant improvements in audio-visual classification on AudioSet-20K (+**2.5** mAP and +**3.4** mAP), AudioSet-2M (+**2** mAP and +**4.1** mAP), and VGGSound (+**2.2%** and +**3.1%**), respectively. It is worth pointing out that training the largest variant of our model, AVSiam-Huge, is still a lot cheaper than training any of the AV-MAE, MAViL-Stage1, MAViL-Stage2 baselines (i.e., **800** vs. **2854**, **1280**, and **5120** V100 hours). Moreover, the improvements with our model from Base to Large are greater than those seen in AV-MAE [20], with gains on AudioSet-2M (+**2 mAP** vs. +**1.8 mAP**) and VGGSound (+**2.2%** vs. +**0.8%**). These results demonstrate that AVSiam scales well both with respect to the training dataset size as well as the model size. We also note that our largest variant, AVSiam-Huge, achieves the state-of-the-art results on AudioSet-20K [19] (**45.0** mAP), AudioSet-2M [19] (**54.1** mAP), and VGGSound [11] (**68.0%** accuracy) while also requiring only **15%** of pretraining time of the previous best-performing method, MAViL-Stage2.

5.2 Audio-visual Retrieval Results

In Table 2, we compare AVSiam with two variants of the CAV-MAE [26] on video-to-audio and audio-to-video retrieval on AudioSet, VGGSound and MSR-VTT. We note that the CAV-MAE$^+$ baseline is trained with a larger batch size

Table 3: Evaluations of AVSiam against MAViL [28] after both models were finetuned on MSR-VTT [76] for audio-to-video retrieval (**left**) and comparisons in throughput performance (**right**). All models are implemented using the same GPU hardware.

(a) A2V Retrieval on MSR-VTT.

Method	MAViL	AVSiam-Base
R@1 Accuracy	22.8	**24.3**

(b) Throughput Comparison.

Method	MAViL	CAV-MAE	AVSiam
Samples/Sec. ↑	3.84	22.5	**75.4**

than the standard CAV-MAE variant. Our results indicate that for video-to-audio retrieval, AVSiam-Base outperforms both CAV-MAE variants on AudioSet (i.e., **19.7** vs. **18.8**), VGGSound (i.e., **19.0** vs. **14.8**) and MSR-VTT (i.e., **9.3** vs. **7.6**) while requiring significantly fewer parameters (i.e., **100M** vs. **164M**). We also observe that the performance gap of our model w.r.t CAV-MAE variants is even larger for audio-to-video retrieval where AVSiam-Base outperforms CAV-MAE$^+$ in AudioSet (R@1: **17.6** vs. **15.1**), VGGSound (R@1: **20.4** vs. **12.8**) and MSR-VTT (R@1: **16.1** vs. **13.3**) Based on these results, we hypothesize that using a single shared audio-visual backbone is beneficial for audio-vsual retrieval tasks compared to using separate audio and visual encoders. This is because a shared encoder projects audio and visual inputs into a common latent representation space, which may be beneficial for retrieving the correct videos based on the audio inputs and vice-versa. We also note that the performance gap between AVSiam and CAV-MAE$^+$ is bigger on VGGSound than on AudioSet and MSR-VTT. This is because VGGSound contains higher-quality audio-visual pairs, which makes it more suitable for evaluating video-to-audio retrieval. Overall, despite using significantly fewer parameters, AVSiam outperforms all competing approaches by a large margin on all audio-visual retrieval benchmarks.

Additional Results on MSR-VTT. In addition to our zero-shot audio-visual retrieval results, in Table 3(a), we also include the results after finetuning on MSR-VTT as was done in [28]. We then compare our AVSiam-Base variant with the state-of-the-art MAViL baseline. Note that we follow the evaluation protocol of MaViL [28], and only report audio-to-video retrieval results using the R@1 metric. Based on these results, we observe that AVSiam-Base achieves better audio-to-video retrieval results than MAViL (**24.3** vs. **22.8**).

5.3 Throughput Comparison

In Table 3(b), we also include throughput comparisons of our method and the two best performing audio-visual approaches, CAV-MAE and MaViL. The throughput is measured using the number of processed samples per second. All methods are implemented using identical GPU hardware (i.e., single NVIDIA A5000). Our results indicate that AVSiam-Base achieves the highest samples per second throughput (**75.4**), outperforming both CAV-MAE (**22.5**) and MAViL-Stage2 (**3.84**). This suggests that our shared audio-visual encoder design is helpful for improving not only the parameter efficiency but also its throughput.

Table 4: Multi-ratio Masking vs. Fixed-ratio Masking. We compare our multi-ratio input masking scheme with various fixed-ratio masking schemes and the contrastive learning baseline. All methods, including all of our multi-ratio masking variants, are trained only with a contrastive objective (no MAE loss). We use audio-visual classification accuracy and efficiency in terms of V100 GPU pretraining hours as our metrics of comparison. All variants are pretrained on AudioSet-2M and finetuned on AudioSet-20K.

	AS-20K (mAP)↑					
A Ratio \ V Ratio	Acc.↑			V100 Hours ↓		
	25%	50%	75%	25%	50%	75%
25%	40.8	40.3	40.0	362	326	300
50%	39.8	39.5	39.4	163	142	136
75%	39.0	38.8	38.6	138	132	**120**
Contrastive Learning	40.4			510		
Multi-ratio Masking (Ours)	**41.3**			160		

Table 5: Comparison with a Separate-encoder Baseline. We compare our AVSiam with the AVSep baseline that uses separate audio and visual encoders on AudioSet-2M. AVSiam achieves similar or even better performance than AVSep while being significantly more parameter-efficient. Compared to the AVSep baseline, our approach also requires **1.7**× and **1.9**× less GPU memory on Base and Large variants, respectively.

Method	#Param ↓	A	V	A+V	GPU Mem.
AVSep-Base	186M	**45.7**	27.3	49.5	7.0G
AVSiam-Base	**100M**	45.2	**30.8**	**50.1**	**4.1G**
AVSep-Large	640M	**48.0**	29.8	52.0	20.6G
AVSiam-Large	**332M**	47.8	**30.1**	**52.1**	**10.9G**

5.4 Ablation Studies

Next, we present our ablation studies to (1) validate the effectiveness of our proposed multi-ratio scheme, (2) compare our model with an equivalent separate-encoder baseline, (3) study the effect of using audio vs visual encoder as our shared backbone, (4) assess the generalization of our model to audio-only and video-only settings, and lastly, (5) investigate the importance of unsupervised audio-visual pretraining objectives.

The Effectiveness of a Multi-Ratio Scheme. In Table 4, we compare the fixed ratio masking baselines with our multi-ratio scheme using two metrics: (1) mAP audio-visual classification accuracy on AudioSet-20K and (2) the pretraining cost on AudioSet-2M (in V100 GPU hours). All variants in the Table 4 are first pretrained on AudioSet-2M and then finetuned on AudioSet-20K. All the methods in Table 4 are trained only with a contrastive objective (no MAE loss). Based on these results, we first note that the multi-ratio masking scheme variant outperforms all fixed-ratio masking variants. In particular, our multi-ratio mask-

Table 6: Comparing ViT vs. AST encoders. We compare our default method that uses ViT as its shared audio-visual encoder with a variant that uses an Audio Spectrogram Transforme AST [23] instead of a ViT. We report that using ViT as our shared audio-visual encoder produces significantly better results in visual-only and audio-visual classification settings.

	AS-20K (mAP↑)			VGGSound (Acc.↑)		
Method	A	V	A+V	A	V	A+V
AVSiam-Base (w/ AST)	**37.5**	9.6	38.1	**61.3**	30.4	62.5
AVSiam-Base (w/ ViT)	36.5	**23.7**	**41.6**	55.7	**46.0**	**64.9**

Table 7: Generalization to Missing Audio and Missing Video Inputs. We compare the performance of AVSiam and CAV-MAE in missing audio and video input settings. Based on these results, we observe that AVSiam can handle missing modalities robustly.

	AS-2M (mAP↑)		VGGSound (Acc.↑)	
Method	Missing A.	Missing V.	Missing A.	Missing V.
CAV-MAE	11.1	43.3	27.3	51.8
AVSiam-Base	**30.8**	**45.2**	**46.0**	**55.7**

ing outperforms the best fixed-ratio masking scheme (**41.3** mAP vs. **40.8** mAP) while being significantly cheaper to pretrain (**160** vs. **362** V100 GPU hours). Furthermore, our multi-ratio masking approach also outperforms the standard contrastive learning baseline (**41.3** vs **40.4** mAP) while also being cheaper to pretrain (**160** vs. **510** V100 GPU hours). Next, we observe that although the most efficient fixed-ratio masking scheme needs fewer pretraining hours (i.e., **120** and **160** V100 hours), it also performs much worse than our multi-ratio masking scheme (i.e., **38.6** mAP and **41.3** mAP). Thus, these results suggest that our multi-ratio input masking improves performance while being more efficient than the best-performing fixed-ratio masking or contrastive learning baselines.

Comparison with a Separate-encoder Baseline. In Table 5, we also compare our AVSiam model with a baseline that uses separate audio and visual encoders (referred to as AVSep). Both baselines use the same implementation, except that our model uses a shared audio-visual encoder, whereas AVSep uses separate audio and visual encoders. Our results indicate that despite a significantly smaller model capacity (**1.86**× and **1.92**× fewer parameters on Base and Large variant respectively), our AVSiam achieves comparable or even slightly better performance than the variant with separate audio and visual backbones on AudioSet-2M. Our model also requires significantly less GPU memory (**4.1G** vs. **7.0G** on Base and **10.9G** vs. **20.6G** on Large).

Using ViT vs. AST as a Shared Encoder. In Table 6, we also compare our default approach that uses ViT as its shared audio-visual encoder with a variant that uses an Audio Spectrogram Transformer (AST) [23] instead of a ViT. Based on these results, we observe that the AST-based variant achieves

Table 8: Importance of Pretraining Objectives. We analyze the impact of various pretraining objectives on downstream audio-visual classification performance. We observe that pretraining with a joint contrastive and reconstruction loss ($\mathcal{L}_c + \mathcal{L}_{rec}$) achieves the best results.

Method	V100 Hours	AS-20K	VGGSound
No Self-Supervised Pretraining	-	39.9	61.2
AV Contrastive Learning ($\mathcal{L}_c$)	**160**	41.3	64.4
AVSiam-Base ($\mathcal{L}_c + \mathcal{L}_{rec}$)	177	**41.6**	**64.9**

better results for audio modeling (i.e., +**1** mAP on AudioSet-20K and +**5.6%** on VGGSound) but significantly worse results for visual and audio-visual classification on AudioSet-20K (**-14.1** and **-3.5** mAP) and VGGSound (**-15.6**% and **-2.4**%). Thus, we use ViT as our shared audio-visual backbone.

Generalization to Missing Audio and Missing Video Inputs. In Table 7, we evaluate our AVSiam model and CAV-MAE on missing audio (visual-only) and missing video (audio-only) inputs. Our results suggest that our model produces significantly better results than CAV-MAE in these settings. In particular, in the missing video setting, our model outperforms CAV-MAE on AudioSet-2M (i.e., **45.2** mAP vs. **43.3** mAP) and VGGSound (i.e., **55.7%** vs. **51.8%**). Furthermore, in the missing audio input setting, the performance gap between CAV-MAE and AVSiamis even larger on both AudioSet-2M (i.e., **30.8** mAP vs. **11.1** mAP), and VGGSound (i.e., **46.0%** vs. **27.3%**). These results show the flexibility of our model to robustly handle missing modalities (e.g., missing audio and missing video inputs). [**YB:**updated]

Importance of Pretraining Objectives. Lastly, we examine the significance of our pretraining objectives. In Table 8, we report the downstream audio-visual classification accuracy on the AudioSet-20K and VGGSound datasets. We observe that directly performing supervised fine-tuning (i.e., no unsupervised pretraining) leads to significantly worse performance on both AudioSet-20K (**-1.7** mAP) and VGGSound (**-3.7%**), compared to our default variant that uses unsupervised pretraining. Pretraining with an audio-visual contrastive objective only ($\mathcal{L}_c$) leads to a boost in performance (i.e., +**1.4** mAP on AudioSet-20K and +**3.2%** on VGGSound). Furthermore, incorporating the MAE objective in the multimodal decoder ($\mathcal{L}_{rec}$) produces a slight improvement (+**0.3** mAP and +**0.5%**) with a small increase in the computational cost (+**17** V100 hours).

5.5 Qualitative Results

Lastly, in Figure 4, we visualize audio and visual features embeddings extracted using (1) a baseline that uses separate audio and visual encoders (i.e., CAV-MAE [26]) and (2) our shared-weight encoder approach (i.e., AVSiam-Base). The visualization is done using t-SNE [67] on the VGGSound dataset. Each point in the plot represents a single input (+ for audio and • for visual), while different colors depict distinct audio-visual categories. Based on this visualization, we first

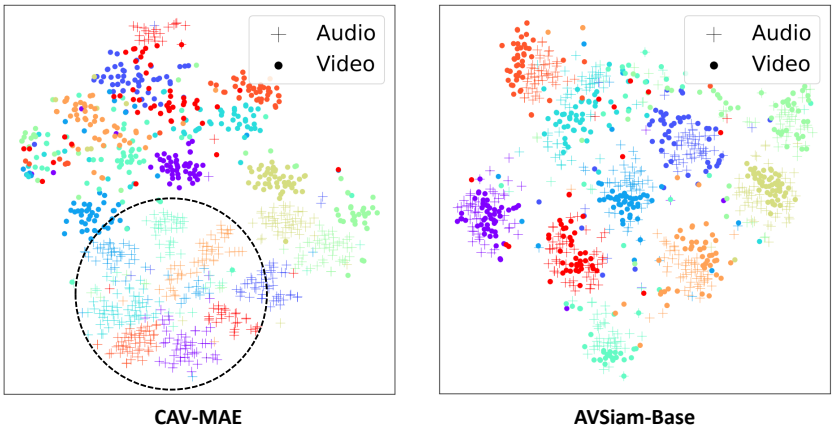

Fig. 4: t-SNE Audio and Image Embedding Visualization. We use t-SNE to visualize the audio and visual features extracted by (1) a baseline that uses separate audio and visual encoders (i.e., CAV-MAE) and (2) our shared-weight encoder method (i.e., AVSiam-Base) on the VGGSound dataset. Each point in the plot represents a single input (+ for audio and ● for visual), while different colors depict distinct audio-visual categories. Based on this illustration, we observe that AVSiam learns more semantically separable features than CAV-MAE. Furthermore, unlike CAV-MAE, AVSiam groups audio and visual features corresponding to the same audio-visual category into the same clusters. This suggests that compared to the methods that use separate audio and visual encoders, our AVSiam with a shared-weight audio-visual encoder learns to encode audio and visual features into a more similar latent space.

observe that our AVSiam produces semantically more separable features than CAV-MAE. Additionally, we observe that, unlike CAV-MAE, AVSiam groups audio and visual features corresponding to the same audio-visual category into the same clusters. This suggests that compared to the approaches that use separate audio and visual encoders, our method encodes audio and visual features into a more similar latent space, which is particularly helpful for tasks such as audio-visual retrieval, as demonstrated by our quantitative results above.

6 Conclusions

In this paper, we propose AVSiam, a framework that uses a single shared ViT encoder for audio and visual data and a novel multi-ratio masking scheme for efficient audio-visual pretraining. AVSiam is fast, scalable, and memory-efficient, and it achieves state-of-the-art results on multiple audio-visual classification and retrieval datasets. In the future, we plan to leverage the efficiency and scalability of our AVSiam model and scale it to even larger datasets and model sizes. We will also extend our framework to other audio-visual understanding tasks, such as audio-visual question-answering, event localization, and segmentation.

Acknowledgments

We thank Feng Cheng, Md Mohaiminul Islam, Ce Zhang, Yue Yang, and Soumitri Chattopadhyay for their helpful discussions. This work was supported by the Sony Faculty Innovation Award, Laboratory for Analytic Sciences via NC State University, ONR Award N00014-23-1-2356.

References

1. Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In *NeurIPS*, 2021. 3, 4
2. Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *NeurIPS*, 2020. 3
3. Humam Alwassel, Dhruv Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 3
4. Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 3
5. Relja Arandjelović and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 3
6. Yuki M Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, 2020. 3
7. Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *NeurIPS*, 2016. 3
8. Alan Baade, Puyuan Peng, and David Harwath. Mae-ast: Masked autoencoding audio spectrogram transformer. In *INTEERSPEECH*, 2022. 3
9. Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022. 3
10. Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *ICML*, 2022. 3
11. Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 3, 6, 7, 9
12. Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2023. 1
13. Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *CVPR*, 2021. 3
14. Haoyue Cheng, Zhaoyang Liu, Hang Zhou, Chen Qian, Wayne Wu, and Limin Wang. Joint-modal label denoising for weakly-supervised audio-visual video parsing. In *ECCV*, 2022. 2, 3
15. Yong Dai, Duyu Tang, Liangxin Liu, Minghuan Tan, Cong Zhou, Jingquan Wang, Zhangyin Feng, Fan Zhang, Xueyu Hu, and Shuming Shi. One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code. *arXiv Preprint*, 2022. 4

16. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 4, 5
17. Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 1
18. Haytham M Fayek and Anurag Kumar. Large scale audiovisual learning of sounds with weakly labeled data. In *IJCAI*, 2020. 8
19. Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, 2017. 3, 6, 7, 9
20. Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *ICCV*, 2023. 2, 3, 4, 6, 8, 9
21. Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *CVPR*, 2023. 3
22. Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *CVPR*, 2022. 3
23. Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *INTEERSPEECH*, 2021. 12
24. Yuan Gong, Yu-An Chung, and James Glass. Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. *TASLP*, 2021. 2
25. Yuan Gong, Alexander H Liu, Andrew Rouditchenko, and James Glass. Uavm: A unified model for audio-visual learning. *IEEE Signal Processing Letters*, 2022. 3
26. Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. In *ICLR*, 2023. 2, 3, 4, 5, 6, 7, 8, 9, 13
27. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 5, 6
28. Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Haoqi Fan, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, and Christoph Feichtenhofer. Mavil: Masked audio-video learners. In *NeurIPS*, 2023. 1, 2, 3, 4, 5, 6, 7, 8, 10
29. Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. In *ICLR*, 2022. 3
30. Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *ICML*, 2021. 3, 8
31. Simon Jenni, Alexander Black, and John Collomosse. Audio-visual contrastive learning with temporal self-supervision. In *AAAI*, 2023. 3
32. Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *ICLR*, 2021. 1
33. Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *ICCV*, 2021. 8, 9
34. Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, 2022. 1

35. Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, 2023. 2, 5
36. Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, et al. Mst: Masked self-supervised transformer for visual representation. In *NeurIPS*, 2021. 5
37. Valerii Likhosherstov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv Preprint*, 2021. 3, 4
38. Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. In *ECCV*, 2022. 2, 3
39. Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *CVPR*, 2023. 2
40. Yan-Bo Lin, Hung-Yu Tseng, Hsin-Ying Lee, Yen-Yu Lin, and Ming-Hsuan Yang. Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. In *NeurIPS*, 2021. 3
41. Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *ACCV*, 2020. 3
42. Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. *arXiv Preprint*, 2022. 3
43. Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. In *ICLR*, 2021. 3
44. Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Contrastive learning of global and local audio-visual representations. In *NeurIPS*, 2021. 3
45. Tanvir Mahmud and Diana Marculescu. Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization. In *WACV*, 2023. 3
46. Shentong Mo and Pedro Morgado. A unified audio-visual learning framework for localization, separation, and recognition. In *ICML*, 2023. 2, 4
47. Shentong Mo, Weiguo Pian, and Yapeng Tian. Class-incremental grouping network for continual audio-visual learning. In *ICCV*, 2023. 3
48. Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *NeurIPS*, 2022. 3
49. Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *CVPR*, 2021. 3
50. Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *CVPR*, 2021. 3
51. Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. In *NeurIPS*, 2022. 4
52. Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021. 1, 3, 4, 5, 8
53. Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv Preprint*, 2018. 6
54. Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *ECCV*, 2018. 3
55. Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *ECCV*, 2016. 3
56. Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *ICASSP*, 2020. 3
57. Janani Ramaswamy and Sukhendu Das. See the sound, hear the pixels. In *WACV*, 2020. 3

58. Varshanth Rao, Md Ibrahim Khalil, Haoda Li, Peng Dai, and Juwei Lu. Dual perspective network for audio-visual event localization. In *ECCV*, 2022. 3
59. Adrià Recasens, Jason Lin, João Carreira, Drew Jaegle, Luyu Wang, Jean-baptiste Alayrac, Pauline Luc, Antoine Miech, Lucas Smaira, Ross Hemsley, et al. Zorro: the masked multimodal transformer. *arXiv Preprint*, 2023. 3, 8
60. Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *ICLR*, 2022. 3
61. Nina Shvetsova, Brian Chen, Andrew Rouditchenko, Samuel Thomas, Brian Kingsbury, Rogerio Feris, David Harwath, James Glass, and Hilde Kuehne. Everything at once–multi-modal fusion transformer for video retrieval. In *CVPR*, 2022. 4
62. Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal. Tvlt: Textless vision-language transformer. In *NeurIPS*, 2022. 2, 3
63. Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton Van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *CVPR*, 2022. 2
64. Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *ECCV*, 2020. 1, 3
65. Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 1, 3
66. Michael Tschannen, Basil Mustafa, and Neil Houlsby. Clippo: Image-and-language understanding from pixels only. In *CVPR*, 2023. 3
67. Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 13
68. Tan Wad, Qianru Sun, Sugiri Pranata, Karlekar Jayashree, and Hanwang Zhang. Equivariance and invariance inductive bias for learning from insufficient data. In *ECCV*, 2022. 2
69. Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv Preprint*, 2023. 1, 3
70. Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020. 8
71. Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 5
72. Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *CVPR*, 2021. 1, 3
73. Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *ICCV*, 2019. 3
74. Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *CVPR*, 2022. 1
75. Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, Christoph Feichtenhofer, et al. Masked autoencoders that listen. In *NeurIPS*, 2022. 3
76. Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 3, 7, 10
77. Hanyu Xuan, Zhenyu Zhang, Shuo Chen, Jian Yang, and Yan Yan. Cross-modal attention network for temporal inconsistent audio-visual event localization. In *AAAI*, 2020. 3
78. Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *ECCV*, 2022. 4

79. Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *TMLR*, 2022. 1
80. Yunhua Zhang, Hazel Doughty, Ling Shao, and Cees GM Snoek. Audio-adaptive activity recognition across video domains. In *CVPR*, 2022. 1
81. Yuanyi Zhong, Haoran Tang, Jun-Kun Chen, and Yu-Xiong Wang. Contrastive learning relies more on spatial inductive bias than supervised learning: An empirical study. In *ICCV*, 2023. 2