Unveiling the Power of Audio-Visual Early Fusion Transformers with Dense Interactions through Masked Modeling

Shentong Mo
Carnegie Mellon University
shentong.mo@sv.cmu.edu

Pedro Morgado
University of Wisconsin-Madison
pmorgado@wisc.edu

https://github.com/stoneMo/DeepAVFusion

Abstract

Humans possess a remarkable ability to integrate auditory and visual information, enabling a deeper understanding of the surrounding environment. This early fusion of audio and visual cues, demonstrated through cognitive psychology and neuroscience research, offers promising potential for developing multimodal perception models. However, training early fusion architectures poses significant challenges, as the increased model expressivity requires robust learning frameworks to harness their enhanced capabilities. In this paper, we address this challenge by leveraging the masked reconstruction framework, previously successful in unimodal settings, to train audio-visual encoders with early fusion. Additionally, we propose an attention-based fusion module that captures interactions between local audio and visual representations, enhancing the model's ability to capture fine-grained interactions. While effective, this procedure can become computationally intractable, as the number of local representations increases. Thus, to address the computational complexity, we propose an alternative procedure that factorizes the local representations before representing audio-visual interactions. Extensive evaluations on a variety of datasets demonstrate the superiority of our approach in audio-event classification, visual sound localization, sound separation, and audio-visual segmentation. These contributions enable the efficient training of deeply integrated audio-visual models and significantly advance the usefulness of early fusion architectures.

1. Introduction

Humans naturally integrate audio-visual information to perceive and understand the environment. Several studies in cognitive psychology and neuroscience, as well as classic perceptual illusions such as the McGurk effect, demonstrate that such audio-visual fusion can occur early on in the perceptual stack, enabling deeper integration of the two modalities. Early fusion models aim to emulate this human-like perception. Specifically, it refers to the process of integrating auditory and visual cues *at an early stage* of a multi-modal perception model, in order to leverage the synergistic effects of both modalities. This can be especially powerful if the fusion process can attend to and establish connections between local components of the visual and audio signal (*e.g.*, the frequencies of someone's voice with the pixels of their lips). In sum, early fusion of local audio-visual interactions holds the promise of a deeper and more sophisticated understanding of audio-visual content, critical for many real-world applications such as visually guided source separation, localization/segmentation, and multi-modal recognition.

Despite the potential, audio-visual early fusion architectures have remained under-explored, with the majority of the recent literature focusing on late fusion or no fusion at all. One recent prominent research direction [1, 47] seeks to learn independent uni-modal encoders through contrastive learning by requiring the representations of associated audio-visual pairs to be synchronized in latent space. However, contrastive learning is not compatible with early fusion, as the connections between uni-modal representations lead to trivial solutions to the contrastive learning problem. Another line of recent work [20, 27] learns encoders with late audio-visual fusion through a combination of contrastive learning and masked reconstruction, focusing on downstream tasks that require shallow integration of audio-visual features, such as audio-visual action recognition.

One potential reason is that early fusion architectures pose a significant training challenge due to the substantial increase in model expressivity it offers. As a result, robust learning frameworks are required to ensure the creation of high-quality models and effectively harness their enhanced capabilities. In this work, we demonstrate the effectiveness of a masked reconstruction framework for training audio-

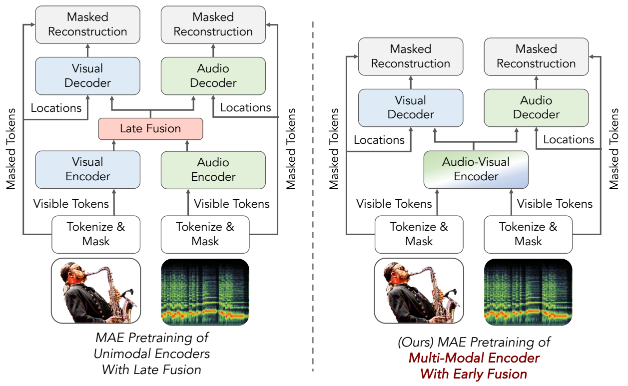

Figure 1. Unlike prior works on Audio-Visual MAEs which learn either late fusion encoders or separate encoders altogether, we demonstrate that masked reconstruction is especially useful for learning deeply integrated audio-visual encoders with powerful early fusion modules.

visual transformers with early-fusion of local interactions (Fig. 1). Our hypothesis is that the simultaneous reconstruction of audio and visual inputs from a limited audio-visual context promotes the learning of deeply integrated representations, as the model is required to understand the fine interactions between the two modalities. Furthermore, masked reconstruction has already been successfully deployed in training uni-modal representations for a variety of input signals, including text [11], images [3, 22], and audio [28], as well as for learning late fusion representations [20, 27]. However, unlike what was previously observed with uni-modal or late fusion models, our work shows that masked audio-visual reconstruction with early fusion encoders results in an interesting emergent property. While uni-modal representations encode the low-level details required for masked reconstruction, we observed that high-level semantics (surprisingly) emerged from the tokens used for audio-visual fusion.

In addition to demonstrating the effectiveness of masked reconstruction for early fusion transformers, we propose a novel attention-based fusion module, which can effectively attend to the *interactions between local audio and visual representations*, thereby enhancing its capacity to capture localized interactions between the two modalities. Take, for instance, the sound of a dog barking. The sound signals the presence of a dog and offers clues about its state, such as an open mouth. While the link between the dog and its bark can be made at a global level, our fusion module allows the model to delve deeper into local interactions, like the correlation between the dog's mouth and the barking sound's spectral frequencies. As a result, localized fusion boosts the quality of learned representations for numerous tasks. However, it necessitates the model to explicitly represent all pairwise interactions of local audio and visual tokens. Given the vast number of tokens that current transformers process, dense local interactions can quickly become unmanageable. To enhance efficiency, our fusion module factorizes uni-modal representations, thereby circumventing the need to model all pairwise interactions.

In sum, we propose an attention-based fusion module that can attend to local audio-visual interactions and deploy it to learn deeply integrated audio-visual representations by fusing uni-modal representations at early layers in the architecture. We also demonstrate that when paired with masked reconstruction, our framework, denoted DeepAVFusion, can learn strong audio-visual representations, yielding state-of-the-art performance on a variety of audio-visual tasks, including visually guided source separation, localization, and segmentation. We extensively evaluated DeepAVFusion on a variety of audio-visual datasets, and conducted thorough ablation studies, showing the importance of early fusion, dense interactions, and uni-modal pre-training when learning deeply integrated audio-visual representations.

2. Related Work

Audio-Visual Representations Learning. Audio-visual representations learning has been addressed in many previous works [1, 2, 14, 16, 23, 24, 30, 40, 43, 45–47, 50, 51, 65, 66] to learn the audio-visual correlation between two distinct modalities from videos. Such cross-modal alignments are beneficial for many audio-visual tasks, such as audio-event localization [31, 32, 37, 41, 42, 53, 62], audio-visual spatialization [4, 17, 44, 45], audio-visual navigation [4–6] and audio-visual parsing [33, 39, 54, 61]. In this work, our main focus is to learn transferrable audio-visual representations from masked audio-visual reconstruction, which is more challenging than the above-mentioned tasks.

Masked Representations Learning. Masked representation learning aims to learn self-supervised representations by reconstructing desired features of masked data given unmasked parts as clues. In the recent years, masked representation learning has achieved promising results in natural language processing [9, 11, 35, 52, 59] and computer vision [3, 8, 12, 15, 22, 58, 60, 63] community. Typically, BERT [11] randomly masked 15% of word tokens in the sentence and recovered them with unmasked words to learn generalizable textual features via a self-attention transformer [56]. A block-wise masking strategy was proposed in BEiT [3] to reconstruct discrete tokens of masked image patches for pre-training transferrable visual representations. To simplify the masked image encoding framework, MAE [22] directly reconstructed missing pixels of 75% masked patches using vision transformers [13] for self-supervised pre-training.

Audio-visual Masked Autoencoders. More recently, researchers introduced diverse masking pipelines [19, 20, 27] to show the effectiveness of masked modeling in learning audio-visual representations. For example, CAV-MAE [20] combined contrastive learning and masked modeling to capture a joint and coordinated audio-visual representation.

They tried to add a joint encoder on audio-visual features from the last attention block of single-modality encoders to fuse cross-modal information for audio-visual contrastive objectives. MAViL [27] extended masked audio-video reconstruction with masked intra- and inter-modal contrastive learning and self-training by recovering joint audio-video contextualized representations. Despite their promising performance, they ignored the importance of the early fusion of audio-visual features in masked audio-visual reconstruction. In contrast to AV-MAE [19] that simply concatenated the audio and visual tokens before passing them through the joint transformer for early fusion, we will design an early fusion module with interactions between local and visual representations for audio-visual masked auto-encoders. Our audio-visual interactions are different from fusion bottlenecks in MBT [48] that forces information between different modalities to pass through a small number of bottleneck latent. However, we develop a fully novel attention-based fusion module that can effectively attend to interactions between local audio and visual representations, thus enhancing its ability to capture fine-grained interactions between the two modalities.

Audio-visual Early Fusion. Audio-visual fusion has been the topic of extensive research, with several works proposing a variety of architectures to aggregate multi-modal representations [19, 48, 49]. An example of an early fusion architecture was proposed by Owens *et al.* [49] to learn representations from audio-visual correspondences, by concatenating features from small unimodal encoders and feeding them into a fused audio-visual network for the early fusion of multisensory features. Beyond this early work, most recent papers on audio-visual fusion [19, 48] propose alternative fusion mechanisms, either through weight sharing [19] or token-based fusion [48]. However, since all these works focused on downstream tasks (mostly classification) that do not require fine multi-modal understanding, the benefits of early fusion were not evident, with mid-level fusion often achieving optimal performance. Different from them, we proposed an early fusion model that is designed for fine multi-modal understanding, by attending to local interactions between the audio and the visual while performing early fusion. The new architecture design paired with masked reconstruction objective is shown to outperform many of the prior works on a variety of downstream multi-modal applications beyond classification.

3. Method

We aim to learn audio-visual representations that can be effectively transferred to a variety of downstream tasks that require a detailed understanding of audio-visual interactions. To accomplish this, we introduce a novel audio-visual transformer architecture, named DeepAVFusion, that enables the early fusion of audio and visual tokens through dense local interactions. We also show how to effectively train such early fusion transformers using multi-modal masked reconstruction. In this section, we first describe the proposed early fusion transformer and then introduce the multi-modal masked reconstruction framework used for self-supervised pre-training.

3.1. A Transformer Architecture for Joint Audio-Visual Encoding

Downstream tasks such as visually guided sound source separation, localization, and segmentation require a deep understanding of fine audio-visual interactions present in the data. However, late fusion often lacks the expressive power to represent such interactions. We introduce a transformer architecture that enables the early fusion of multi-modal representations through factorized local interactions between audio and visual tokens.

Early fusion architecture While we aim to learn deeply integrated audio-visual representations, foundational uni-modal transformers can still provide a strong starting point for multi-modal learning. As such, we introduce a modular architecture to repurpose existing pre-trained uni-modal transformers. Specifically, we propose a three-branch architecture, illustrated in Fig. 2. The first two branches are uni-modal transformers that process audio and visual tokens, respectively, while the third branch is a multi-modal transformer that updates a set of learnable fusion tokens to fuse audio-visual information. Formally, let $\boldsymbol{X}_0^f \in \Re^{F\times D}$ be F learnable fusion tokens. We progressively update these tokens, at each layer l, through a fusion block Φ_l^f, which can interact with and aggregate information from modality-specific features $\boldsymbol{X}_{l-1}^v, \boldsymbol{X}_{l-1}^a$. Furthermore, fusion tokens $\boldsymbol{X}_{l-1}^f$ are fed into modality-specific blocks, Φ_l^v and Φ_l^a, to modulate uni-modal representations from early stages in the network. Overall, representations are updated as follows

$$\boldsymbol{X}_l^f = \Phi_l^f(\boldsymbol{X}_{l-1}^f, \boldsymbol{X}_{l-1}^v, \boldsymbol{X}_{l-1}^a) \quad (1)$$

$$\boldsymbol{X}_l^v = \Phi_l^v(\boldsymbol{X}_{l-1}^v, \boldsymbol{X}_{l-1}^f) \quad (2)$$

$$\boldsymbol{X}_l^a = \Phi_l^a(\boldsymbol{X}_{l-1}^a, \boldsymbol{X}_{l-1}^f) \quad (3)$$

This token update strategy is repeated for a total of L layers to compute the final representations $\boldsymbol{X}_L^v, \boldsymbol{X}_L^a$ and $\boldsymbol{X}_L^f$.

Modality-specific blocks We use standard transformer self-attention for the modality-specific blocks, Φ^v and Φ^a, while allowing them to further attend to fusion tokens. Unless otherwise specified, we initialize Φ^v and Φ^a from foundational uni-modal models. This is possible since the parameter space of the modality-specific blocks remains unchanged. In particular, we empirically validated the benefits of bootstrapping from [28] and [22], which are pre-trained on large-scale audio and visual datasets, respectively, through a uni-modal masked reconstruction objective.

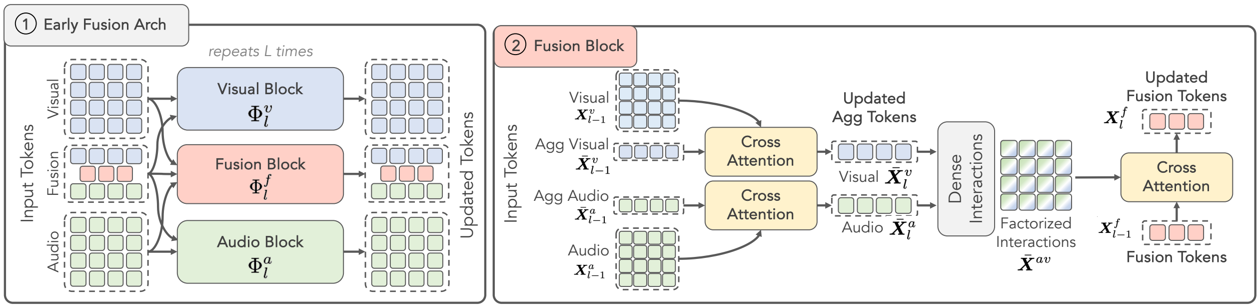

Figure 2. Audio-Visual Early Fusion Transformer with Dense Interactions. The architecture ① is composed of three interconnected branches - the visual, audio, and fusion branches - each with an equal number of transformer blocks. The visual and audio branches process visual and audio tokens, respectively, while simultaneously attending to fusion tokens. The fusion branch updates a set of learnable fusion tokens to fuse audio-visual information. To better attend to local audio-visual interactions, we introduce a new fusion block ②. At each layer l, all patch-level audio (visual) tokens, $\boldsymbol{X}_{l-1}^a(\boldsymbol{X}_{l-1}^v)$, are first aggregated into a small set of aggregation tokens, $\bar{\boldsymbol{X}}_l^a(\bar{\boldsymbol{X}}_l^v)$ by cross-attention. Pairwise audio-visual interactions are then encoded through a linear layer into latents $\bar{\boldsymbol{X}}_l^{av}$, which are used to update the fusion tokens $\boldsymbol{X}_l^f$ by cross-attention.

Fusion blocks The design of the fusion blocks Φ^f is critical to representing fine audio-visual interactions. In the context of transformers, a natural way to combine audio-visual representations is through a standard self-attention block which, in addition to the fusion tokens $\boldsymbol{X}^f$ themselves, it can further attend to all image and audio tokens, $\boldsymbol{X}_{l-1}^v$ and $\boldsymbol{X}_{l-1}^a$. This approach, which we refer to as *token fusion*, has been used in previous works such as MBT [48]. While intuitive, standard self-attention cannot *directly* exploit interactions between local audio-visual representations, which limits their expressivity. To address this limitation, we designed a fusion block that attends to local interactions.

Fusion with Dense Audio-Visual Interactions Interactions between audio and visual data occur between local regions in the image and the audio spectrogram. Given a set of n_a audio tokens $\boldsymbol{X}^a$ and n_v visual tokens $\boldsymbol{X}^v$, aggregation of local multi-modal interactions requires attending to representations from all possible $n_a \times n_v$ pairs, to identify the ones that are most useful for the task at hand. While we experimented with several ways to represent multi-modal interactions, from bilinear aggregation to kernel-based aggregation, we found that linear aggregation was sufficient to capture audio-visual interactions for our tasks. Formally,

$$\boldsymbol{X}_{ij}^{av} = W^a \boldsymbol{X}_i^a + W^v \boldsymbol{X}_j^v \tag{4}$$
$$\boldsymbol{X}^{av} = \left[\boldsymbol{X}_{ij}^{av}; \forall i = 1, \dots, n_a \forall j = 1, \dots, n_v\right]. \tag{5}$$

Fusion tokens $\boldsymbol{X}_l^f$ at layer l are then updated using standard cross-attention [56] over the representation of all local interactions $\boldsymbol{X}^{av}$

$$\boldsymbol{X}_l^f = \mathit{CrossAttention}\left(\boldsymbol{X}_{l-1}^f; \boldsymbol{X}_{l-1}^{av}\right). \tag{6}$$

Factorized Audio-Visual Interactions While the procedure above can attend to local interactions, it requires the model to explicitly represent all possible $n_a \times n_v$ pairs. Given that the number of audio and visual tokens is typically large, the number of interactions can quickly become intractable, resulting in high memory consumption and low throughput. To address this limitation, we introduce modality-specific aggregation tokens updated by aggregation blocks. The goal of the aggregation blocks is to summarize the large number of audio/visual tokens $\boldsymbol{X}^a/\boldsymbol{X}^v$ into a small set of $n_{\bar{a}}/n_{\bar{v}}$ aggregation tokens $\bar{\boldsymbol{X}}^a/\bar{\boldsymbol{X}}^v$, respectively. Specifically, at layer l, aggregation tokens are updated by cross-attention

$$\bar{\boldsymbol{X}}_l^a = \mathit{CrossAttention}\left(\bar{\boldsymbol{X}}_{l-1}^a, \boldsymbol{X}_{l-1}^a\right) \tag{7}$$
$$\bar{\boldsymbol{X}}_l^v = \mathit{CrossAttention}\left(\bar{\boldsymbol{X}}_{l-1}^v, \boldsymbol{X}_{l-1}^v\right). \tag{8}$$

Then, instead of explicitly representing all $n_a \times n_v$ pairwise interactions, we restrict audio-visual interactions to the sets of aggregated tokens. Formally, factorized interactions between aggregated tokens $\bar{\boldsymbol{X}}_i^a$ and $\bar{\boldsymbol{X}}_j^v$ are represented as

$$\bar{\boldsymbol{X}}_{ij}^{av} = W^{\bar{a}} \bar{\boldsymbol{X}}_i^a + W^{\bar{v}} \bar{\boldsymbol{X}}_j^v \tag{9}$$
$$\bar{\boldsymbol{X}}^{av} = \left[\bar{\boldsymbol{X}}_{ij}^{av}; \forall i = 1, \dots, n_{\bar{a}} \forall j = 1, \dots, n_{\bar{v}}\right] \tag{10}$$

and the fusion tokens are updated at layer l by

$$\boldsymbol{X}_l^f = \mathit{CrossAttention}\left(\boldsymbol{X}_{l-1}^f, \bar{\boldsymbol{X}}_l^{av}\right). \tag{11}$$

Since we only represent interactions between the aggregated uni-modal tokens, the number of possible interactions is greatly reduced. We observed that a 700× reduction still maintains the benefits obtained with local dense interactions.

3.2. Learning Early-Fusion Transformers Through Audio-Visual Masked Auto-Encoding

Masked auto-encoders have been shown to learn effective representations for a wide variety of input signals, including images [3, 22] and audio [28]. Formally, the input signal x is partitioned into a set of patch tokens $x = \{x_t\}_{t=1}^T$, which is randomly split into the visible set $X_{\text{vis}} = \{x_t\}_{t \in \mathcal{V}}$ and the masked set $X_{\text{mask}} = \{x_t\}_{t \in \mathcal{M}}$ ($\mathcal{V}$ and $\mathcal{M}$ are non-overlapping index sets). In the case of images, visible and masked tokens, X_{vis} and X_{mask}, are obtained from a grid of small image patches [3, 22], while in the case of audio signals, tokens are obtained either from small windows of a raw waveform or from small 2D patches of the audio spectrogram [28]. Then, to learn representations of the input data, a transformer encoder f is first used to encode the visible tokens $\boldsymbol{X}_{\text{vis}} = f(X_{\text{vis}}) \in \Re^{|\mathcal{V}| \times d}$, where d denotes the dimensionality of the token representation. These representations together with a series of (position marked) mask tokens $\boldsymbol{M}_{\text{mask}}$ are then fed to a decoder $g(\cdot)$ that seeks to reconstruct the masked input tokens $\hat{X}_{\text{mask}} = g(\boldsymbol{X}_{\text{vis}}, \boldsymbol{M}_{\text{mask}})$. To accomplish that, the model is trained to minimize the mean squared reconstruction error

$$\mathcal{L}_{\text{MAE}}\left(X_{\text{mask}}, \hat{X}_{\text{mask}}\right) = \frac{1}{|\mathcal{M}|} \sum_{\substack{x \in X_{\text{mask}} \\ \hat{x} \in \hat{X}_{\text{mask}}}} \|x - \hat{x}\|_2^2 . \tag{12}$$

Following prior work [20, 27], masked reconstruction can be easily extended to multi-modal signals, by simultaneously masking tokens from both audio and visual modalities. In the case of early-fusion transformers, the encoder f is then used to encode the visible tokens, X_{vis}^v and X_{vis}^a,

$$\boldsymbol{X}_{\text{vis}}^v, \boldsymbol{X}_{\text{vis}}^a, \boldsymbol{X}^f = f(X_{\text{vis}}^v, X_{\text{vis}}^a) \tag{13}$$

which, in addition to the uni-modal representations, $\boldsymbol{X}_{\text{vis}}^v$ and $\boldsymbol{X}_{\text{vis}}^a$, also returns a set of fusion representations $\boldsymbol{X}^f$. These representations are then fed to modality-specific decoders to reconstruct both audio and visual masked tokens, $\hat{X}_{\text{mask}}^v = g_v(\boldsymbol{X}_{\text{vis}}^f, \boldsymbol{M}_{\text{mask}}^v)$ and $\hat{X}_{\text{mask}}^a = g_a(\boldsymbol{X}_{\text{vis}}^f, \boldsymbol{M}_{\text{mask}}^a)$, by minimizing the loss

$$\mathcal{L}_{\text{AV-MAE}} = \mathcal{L}_{\text{MAE}}\left(X_{\text{mask}}^v, \hat{X}_{\text{mask}}^v\right) + \mathcal{L}_{\text{MAE}}\left(X_{\text{mask}}^a, \hat{X}_{\text{mask}}^a\right) \tag{14}$$

where $\mathcal{L}_{\text{MAE}}$ is the auto-encoder loss of Eq. 12.

While multi-modal masked reconstruction has been previously explored in [20, 27], these approaches focus on learning separate audio and image encoders, f_a and f_v, with multi-modal representations obtained through a late fusion module f_f. In contrast, we propose to learn a single multi-modal early fusion encoder f, and demonstrate that masked reconstruction is especially useful for *learning multi-modal representations* capable of representing *fine audio-visual interactions* through *early fusion* of *local interactions* between audio and visual tokens.

4. Experiments

We now evaluate the proposed DeepAVFusion on a variety of audio-visual downstream tasks that require a deep understanding of audio-visual interactions.

4.1. Implementation details

Pre-training. For pre-training, we experiment with two datasets, namely VGG-Sounds [7] (with 144k training samples) and AudioSet [18] (with 1.73M training samples), both of them containing diverse videos focused on audio events and sound sources. The audio is represented by log mel spectrograms extracted from $3s$ of audio at a sample rate of 16000Hz. To compute the log spectrograms, we apply an STFT using approximately 50ms windows with a hop size of 15ms, resulting in an input tensor of size 128×196 (128 mel frequency bands over 196 time steps). Since we use image models for the visual component, we randomly extract single frames from within the time window defined by the $3s$ audio snippets. Input frames are augmented by random crops with a minimum area of 0.5 and random horizontal flips, and resized into a 224×224 resolution.

Unless otherwise specified, we initialize the uni-modal transformer blocks using modality-specific encoders pre-trained with uni-modal masked reconstruction objectives [22, 28]. Specifically, we use the ViT-Base model from [22] pre-trained on ImageNet [10], and the Spec-MAE model (with the same ViT-Base architecture) pre-trained on AudioSet [18]. In our base configuration, we add multi-modal fusion blocks on *all* 12 layers of the model, together with 16 multi-modal fusion tokens, 8 visual aggregation tokens, and 8 audio aggregation tokens. For efficiency, we reduced the dimensionality of the MLP embeddings using an MLP ratio of 1 (instead of the standard 4), as well as the dimensionality of space in which the similarity for self and cross-attention is computed to 16 (as opposed to the standard 64).

The models were trained using the Adam optimizer [29]. Ablation and parametric studies were conducted on VGGSound, training for 200 epochs with a total batch size of 512 (split between 2 GPUs and 4 iterations of gradient accumulation). Large-scale training on AudioSet was conducted for 200 epochs with a total batch size of 2048 (split between 16 GPUs and 4 iterations of gradient accumulation). In all cases, we use a base learning rate of $1.5e-4$ adjusted for the effective batch size by the linear scaling rule $lr = blr \times bs/256$ [21], a 40 epoch learning rate warm-up schedule followed by cosine learning rate decay, and a weight decay of 0.05.

Downstream tasks. After pre-training, we fine-tune our models on a variety of downstream tasks.

Visually-guided sound source separation. Following [65], we use a mix-and-separate strategy for training. The pre-trained model is used to obtain audio-visual representations that are fed to a UNet decoder to generate the separated audio. We use a similar separation strategy by predicting a separation mask that is directly applied to the input mixture SFTF. To compute the separation masks, we deploy a 10-layer UNet decoder, which conditions its predictions on (1) the spatially pooled visual representations and (2) five intermediate audio representations obtained during the encoding stage at layers 1, 3, 6, 9, and 12. Following [65], the predicted separation masks are trained to regress the STFT magnitude ratio between the separated source and the mixture at each time-frequency bin. We evaluate on three different datasets: the MUSIC [65] dataset and two subsets from VGG-Sounds [7]. MUSIC [65] consists of 448 untrimmed YouTube music videos of solos and duets from 11 instrument categories, where we use 358 solo videos for training and 90 solo videos for evaluation. This dataset was slightly smaller than the original MUSIC dataset since some videos are no longer publicly available to be downloaded. We also use two subsets of VGG-Sounds [7]: the VGGSound-Instruments subset [25] which includes 32k video clips of 10s lengths from 36 musical instrument classes, and a broader newly defined "VGGSound-Music" subset which includes 40,908 video clips from 49 music categories for training and 1201 clips for testing. As commonly done in the source separation literature, we measure performance using Signal-to-Distortion Ratio (SDR), Signal-to-Artifact Ratio (SAR), and Signal-to-Interference Ratio (SIR).

Audio-visual segmentation. We fine-tune our models on AVSBench [67] for the tasks of Single Source Source Segmentation (S4) and Audio-Visual Segmentation with Semantics (AVSS). Task S4 aims to predict pixel-level segmentation maps of the visible sound sources, while AVSS further requires the model to recognize the class of the sound source (in addition to the segmentation maps). We use the standard train/val/test split [67] and measure performance using mIoU and F1 scores. Similarly to the source separation task, audio-visual representations, obtained from the pre-trained model, are fed to a UNet decoder to generate the segmentation maps. This decoder conditions its predictions on a globally pooled audio representation and intermediate visual feature maps obtained at layers 1, 3, 6, 9, and 12.

Recognition through Linear Probing or Finetuning. We further assess the pre-trained representations on multi-modal recognition tasks by conducting linear probing and fine-tuning evaluations on the VGGSound-Music, VGGSound-All, and the balanced AudioSet dataset. In all cases, we simply attach three linear classifiers on the average pooled visual, audio, and fusion representations, respectively. The three classification logits are then averaged to obtain the final prediction. As usual, on VGGSound datasets, we measure performance by class accuracy, while on AudioSet (a multi-label dataset) we use the average precision (AP) and Area under the ROC curve (AUC) scores averaged across classes.

4.2. Comparison to prior work

To demonstrate the effectiveness of the proposed method, we extensively compare it to prior work on audio-visual downstream tasks that require a detailed understanding of audio-visual data. Unless otherwise specified, we use our model trained on VGGSound for comparisons to prior work, as this enables us to make the fairest comparisons with the largest number of works that require pre-training.

Table 1. Sound source separation performance on the MUSIC and VGGSound datasets.

Method	MUSIC			VGGS-Instruments			VGGS-Music		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
NMF [57]	-0.89	2.38	6.28	-3.52	0.78	6.95	-5.06	0.15	7.03
RPCA [26]	0.78	4.62	7.58	0.36	2.38	7.95	-1.68	1.57	8.26
SoP [65]	3.62	7.95	9.63	2.72	5.67	9.85	1.56	4.59	10.15
MP-Net [64]	3.98	8.36	9.86	3.05	6.12	10.17	1.95	5.02	10.59
CCoL [55]	5.27	8.75	10.52	4.07	6.93	10.85	2.73	5.86	11.03
OneAVM [38]	6.27	10.68	12.35	5.89	7.85	11.23	3.67	6.53	11.85
DeepAVFusion	**7.95**	**12.41**	**12.80**	**6.95**	**9.52**	**13.23**	**5.79**	**8.24**	**13.82**

Table 2. Audio-visual segmentation performance on the AVSBench dataset for the Sound Source Segmentation (S4) and Audio-Visual Segmentation with Semantics (AVSS) tasks.

Method	AVSBench-S4		AVSBench-AVSS	
	mIoU	F1	mIoU	F1
AVS [67, 68]	78.74	87.90	29.77	35.2
LAVISH [34]	80.10	–	–	–
MMVAE [36]	81.74	90.10	–	–
DeepAVFusion	**89.94**	**92.34**	**52.05**	**58.29**

Sound source separation. Table 1 shows the comparison between DeepAVFusion and several prior works, including traditional signal processing methods, NMF [57] and RPCA [26], deep learning methods specialized on the source separation task, Sound-of-Pixels [65] and MP-Net [64], and deep learning methods trained for simultaneous localization and separation, CCoL [55] and OneAVM [38]. While all listed deep learning approaches to source separation use (supervised) ImageNet pre-trained visual encoders, our model shows strong performance using completely unsupervised pre-training. Table 1 shows that DeepAVFusion outperforms prior works on all metrics. For example, on VGGSound-Music, we outperform the recently proposed OneAVM [38] by 2.12 SDR and 1.71 SIR, which uses supervised pre-training for the vision component and uses source separation as just one of multiple pre-training objectives. Although the different models use different architectures for their audio-visual encoders (which make direct comparison challenging to analyze), the improvements demonstrated on this task indicate that visual source separation can benefit from early-fusion to extract rich representations from fine audio-visual interactions.

Audio-Visual Segmentation. We assess the effectiveness of DeepAVFusion for audio-visual segmentation on the AVSBench dataset [67, 68]. We compared the proposed approach to recent state-of-the-art methods (AVS [67], LAVISH [34] and MMVAE [36]) on the S4 and AVSS downstream tasks (sound source segmentation without and with semantics, respectively). We compare against the mIoU and F1 scores shown in the original papers in Table 2. As can be seen, DeepAVFusion outperforms all prior work on both tasks

Audio-visual classification. The above experiments demonstrate the effectiveness of DeepAVFusion on audio-visual

Table 3. Comparing DEEPAVFUSION trained on VGGSound and AudioSet dataset.

Pre-training Dataset	Pre-training Epochs	VGGSound CLS		AudioSet CLS		VGGSS-Music SEP			AVSBench S4	
		LP Acc	Ft Acc	LP AP	Ft AP	SDR	SIR	SAR	mIoU	F1
VGGSound	200	**53.08**	**58.19**	30.25	32.88	5.79	8.24	13.82	89.94	92.34
AudioSet	200	**53.08**	57.91	**32.69**	**34.71**	**6.93**	**9.93**	**13.49**	**90.27**	**92.49**

Table 4. Audio-visual classification performance on VGGSound-Music and VGGSound-All datasets, using linear probing (LP) and finetuning (FT) protocols. Performance is measured in terms of classification accuracy [%].

Method	VGGS-Music		VGGS-All	
	LP	FT	LP	FT
(Image) MAE [22]	48.25	53.18	37.12	43.86
AudioMAE [28]	52.73	58.29	42.05	49.53
CAV-MAE [20]	59.18	67.58	49.83	55.32
DEEPAVFUSION	**65.32**	**72.25**	**53.08**	**58.19**

tasks that require a detailed understanding of audio-visual interactions. We further evaluated the learned representations on recognition tasks. In particular, we compared our model to uni-modal masked reconstruction methods, MAE [22] and AudioMAE [28], as well as to CAV-MAE [20], a late-fusion model trained with both masked reconstruction and contrastive learning objectives. For a fair comparison, we use the same ViT-Base architecture for the uni-modal encoders and tune the model on image-audio pairs obtained from the VGG-Sound dataset.

As can be seen in Table 4, we outperform both uni-modal pre-trained encoders on linear probing and fine-tuning by significant margins. DEEPAVFUSION outperforms the image MAE [22] and AudioMAE [28] by 19.62% and 13.26% on VGGSound-Music and 19.08% and 12.49% on the full VGGSound dataset. These direct comparisons to uni-modal encoders highlight the importance of learning joint representations for multi-modal recognition. Furthermore, our DEEPAVFUSION also significantly outperforms CAV-MAE which optimizes the same audio-visual reconstruction objective of Eq. 14 but does not leverage an early fusion encoder.

It should be noted that multi-modal recognition could be further improved by learning from audio-video data (instead of audio-image pairs). However, video modeling is beyond the scope of this work. The improvements in Table 4 demonstrate nevertheless that even tasks like audio-visual classification of sound sources can benefit from more deeply integrated multi-modal representations.

4.3. Pre-training data

The comparisons above were conducted using VGGSound for pre-training. To study the impact of pre-training data distribution and scale, we also trained our model on the larger AudioSet dataset [18] (with 1.73M training samples). We then tuned the resulting model on several downstream tasks, from classification on both VGGSounds and AudioSet to source separation on VGGSS-Musion and sound source segmentation on AVSBench (S4 task). As can be seen in Table 3, pre-training on AudioSet yields better performance on most downstream tasks, demonstrating the scalability of the proposed method. The only exception is classification on VGGSound, where pre-training on VGGSound yields slightly better performance. This is likely due to a smaller domain gap between pre-training and fine-tuning data.

Table 5. Ablation study of major components of DEEPAVFUSION.

(#)	Unimodal Pre-Training	Fusion	Local AV Interactions	Linear Acc	Sep SDR	Segm mIoU
(1)	✓	Early	✓	**53.08**	**6.53**	**52.05**
(2)	✗	Early	✓	44.36	2.03	29.55
(3)	✓	Early	✗	51.19	6.15	48.71
(4)	✓	Mid	✓	51.21	5.85	48.04
(5)	✓	Late	✓	46.52	5.32	44.32
(6)	✓	None	✗	39.67	4.23	38.32

4.4. Understanding DEEPAVFUSION

The previous section demonstrated the efficacy of early fusion of local audio-visual interactions, with the proposed method outperforming the state-of-the-art on a wide range of audio-visual tasks. In this section, we conduct a thorough analysis of our framework, revealing emergent properties in our model, and assessing the benefits of uni-modal pre-training, fusion depth, and other major design choices of the proposed methodology. To provide a full-picture comparison, we evaluate all models on five downstream tasks, namely linear probing and fine-tuning on VGGSound, sound source separation on VGGSound-Music, sound source localization on Flickr-SoundNet, and audio-visual segmentation with semantics on AVSBench.

Ablation studies. We begin our analysis with a series of ablation studies in Table 5 to measure the impact of three important components of the proposed framework, namely 1) uni-modal pre-training, 2) early fusion, and 3) audio-visual interactions. All models use the same backbone architecture for the uni-modal encoders and are trained to optimize the Audio-Visual MAE objective. Model #1 is the full DEEPAVFUSION framework, and the remaining models relax each of the three components in turn. Model #2 skips uni-modal pre-training, significantly hurting performance when compared to model #1 on all tasks, emphasizing the value of uni-modal pre-training, when learning deeply integrated audio-visual encoders. Model #3 performs early fusion but uses the "token fusion" strategy similar to that used in MBT [48]. Token fusion attends to uni-modal representations and cannot directly capture audio-visual inter-

Table 6. Exploration studies on the efficiency of factorized interactions.

Early Fusion	Factorized Interactions	Speed (video/sec)	Memory (GB)	Linear Acc	Sep SDR	Segm mIoU
✗	✗	**14.5**	**10.3**	46.52	5.32	44.32
✓	✗	4.3	31.6	52.82	6.30	**52.13**
✓	✓	12.8	13.1	**53.08**	**6.53**	52.05

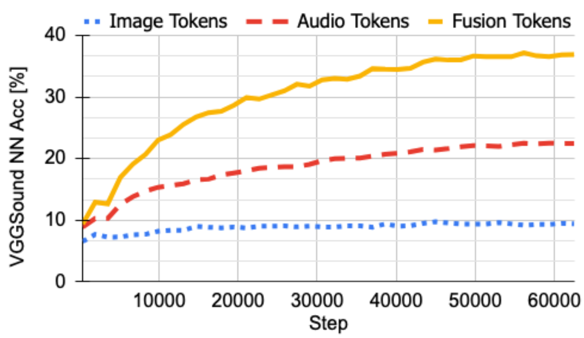

Figure 3. Nearest neighbor accuracy (NNAcc) vs training steps (Step) using Audio, Image, Fusion Tokens from an DEEPAVFUSION trained on VGGSounds.

actions, limiting the expressivity of the fusion block, and the overall performance of the model. Models #4, #5 and #6 use the same fusion modules with dense interactions but perform mid-fusion (at layers 9-12), late-fusion (at layer 12), and no fusion at all. The results indicate that mid-fusion outperforms, late-fusion which in turn is substantially better than no-fusion. Nevertheless, early fusion achieves the best results across all downstream tasks.

Efficiency of factorized interactions. While audio-visual interactions are crucial for the performance of early fusion, they also introduce a significant computational overhead, if not handled properly. Table 6 lists the VRAM consumption and model throughput during inference for three models: a baseline model without fusion, an early fusion model with dense interactions, and our proposed early-fusion model with factorized interactions. Dense interactions enhance the model's effectiveness on subsequent tasks, but at a high computational cost, both in terms of throughput (3.3 times slower) and memory usage (3.1 times higher). Our proposed factorized interactions significantly cut down the computational overhead, while maintaining (and even slightly enhancing) the performance of early fusion with dense interactions.

Emergent semantics. In addition to robust performance across various tasks, the emergent semantics in fusion tokens are worth highlighting. Despite the masked auto-encoder objective typically favoring representations rich in low-level details, fusion tokens surprisingly yield higher-level semantic representations. Fig. 3 shows the nearest neighbor retrieval performance of our DEEPAVFUSION using either the learned visual, audio, or fusion tokens as queries. As can be seen, the retrieval using fusion tokens is substantially better than uni-modal representations (*i.e.* better aligned with the semantics of the dataset). This suggests that fusion tokens aggregate information that is indicative of high-level semantics, while uni-modal representations encode the low-level details required for masked reconstruction.

Table 7. Impact of early fusion and fusion tokens.

#	Linear Acc	Sep SDR	Segm mIoU
12	**53.08**	**6.53**	**52.05**
9	52.76	6.15	49.93
6	52.29	6.12	49.02
3	51.21	5.85	48.04
1	46.52	5.32	44.32
0	39.67	4.23	38.32

(a) # Fusion layers

#	Linear Acc	Sep SDR	Segm mIoU
1	51.86	5.94	50.67
8	51.97	5.83	51.14
16	**53.08**	**6.53**	**52.05**
32	52.18	6.43	50.60

(b) # Fusion tokens

#	Linear Acc	Sep SDR	Segm mIoU
4	51.78	5.68	49.59
8	**53.08**	**6.53**	**52.05**
16	52.25	6.16	50.89

(c) # Aggr tokens

Impact of miscellaneous design choices. Lastly, we examine the effects of the number of fusion layers, fusion tokens and aggregation tokens in Table 7. These results reveal that early fusion is crucial for optimal performance, with model performance improving as fusion depth increases. Furthermore, DEEPAVFUSION can benefit from a relatively large number of tokens, with the performance saturating after 16 fusion tokens and 8 aggregation tokens.

5. Conclusion

In this work, we present DEEPAVFUSION, a simple yet effective early fusion approach with dense interactions that achieves efficient audio-visual pre-training for audio-visual masked auto-encoders. We introduce learnable fusion tokens to aggregate modality-specific information with early fusion from each transformer block of audio and visual encoders, where two variants of parallel and sequential flows are proposed to achieve early fusion between fusion tokens and audio-visual patch tokens. Furthermore, we leverage multi-modal blocks with dense interactions to fuse fusion tokens and patches across audio-visual transformer blocks and achieve efficient downstream fine-tuning with factorized attention blocks. Empirical experiments on Flick-SoundNet, VGG-Instruments, VGG-Music, VGGSound-All, and AVS-Bench datasets demonstrate the state-of-the-art performance of our DEEPAVFUSION in linear probing, fine-tuning classification, visual sound localization, sound separation, and audio-visual segmentation.

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 1, 2

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2022. 2, 4, 5

[4] Changan Chen, Unnat Jain, Carl Schissler, S. V. A. Garí, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 17–36, 2020. 2

[5] Changan Chen, Sagnik Majumder, Al-Halah Ziad, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[6] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022. 2

[7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 5, 6

[8] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 2

[9] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia. Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[12] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for BERT pre-training of vision transformers. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2023. 2

[13] 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 *Proceedings of International Conference on Learning Representations (ICLR)*, 2021. 2

[14] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018. 2

[15] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *Proceedings of Advances In Neural Information Processing Systems (NeurIPS)*, 2022. 2

[16] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10478–10487, 2020. 2

[17] Ruohan Gao and Kristen Grauman. 2.5d visual sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 324–333, 2019. 2

[18] 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 *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 5, 7

[19] 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 (ICCV)*, pages 16144–16154, 2023. 2, 3

[20] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *Proceedings of The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 1, 2, 5, 7

[21] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 5

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 2, 3, 4, 5, 7

[23] John Hershey and Michael Casey. Audio-visual sound separation via hidden markov models. *Advances in Neural Information Processing Systems*, 14, 2001. 2

[24] Di Hu, Feiping Nie, and Xuelong Li. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9248–9257, 2019. 2

[25] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10483–10492, 2022. 6

[26] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 57–60, 2012. 6

[27] 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. *arXiv preprint arXiv:2212.08071*, 2022. 1, 2, 3, 5

[28] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. In *Proceedings of Advances In Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3, 4, 5, 7

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[30] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[31] Yan-Bo Lin and Yu-Chiang Frank Wang. Audiovisual transformer with instance attention for audio-visual event localization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 2

[32] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2002–2006, 2019. 2

[33] 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 *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[34] 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*, 2023. 6

[35] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[36] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 954–965, 2023. 6

[37] Shentong Mo and Pedro Morgado. Benchmarking weakly-supervised audio-visual sound localization. In *European Conference on Computer Vision (ECCV) Workshop*, 2022. 2

[38] Shentong Mo and Pedro Morgado. A unified audio-visual learning framework for localization, separation, and recognition. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. 6

[39] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[40] Shentong Mo and Yapeng Tian. Semantic-aware multi-modal grouping for weakly-supervised audio-visual video parsing. In *European Conference on Computer Vision (ECCV) Workshop*, 2022. 2

[41] Shentong Mo and Yapeng Tian. Audio-visual grouping network for sound localization from mixtures. *arXiv preprint arXiv:2303.17056*, 2023. 2

[42] Shentong Mo and Yapeng Tian. AV-SAM: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*, 2023. 2

[43] Shentong Mo, Jing Shi, and Yapeng Tian. DiffAVA: Personalized text-to-audio generation with visual alignment. *arXiv preprint arXiv:2305.12903*, 2023. 2

[44] Pedro Morgado, Nuno Vasconcelos, Timothy Langlois, and Oliver Wang. Self-supervised generation of spatial audio for 360 video. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[45] Pedro Morgado, Yi Li, and Nuno Vasconcelos. Learning representations from audio-visual spatial alignment. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 4733–4744, 2020. 2

[46] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12934–12945, 2021.

[47] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12475–12486, 2021. 1, 2

[48] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3, 4, 7

[49] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 3

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

[51] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4358–4366, 2018. 2

[52] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*, 2019. 2

[53] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018. 2

[54] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. In *Proceedings of European Conference on Computer Vision (ECCV)*, page 436–454, 2020. 2

[55] Yapeng Tian, Di Hu, and Chenliang Xu. Cyclic co-learning of sounding object visual grounding and sound separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2745–2754, 2021. 6

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, page 5998–6008, 2017. 2, 4

[57] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007. 6

[58] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan L. Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[59] Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 2985–3000, 2023. 2

[60] Jiantao Wu and Shentong Mo. Object-wise masked autoencoders for fast pre-training. *arXiv preprint arXiv:2205.14338*, 2022. 2

[61] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1326–1335, 2021. 2

[62] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6291–6299, 2019. 2

[63] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[64] Xudong Xu, Bo Dai, and Dahua Lin. Recursive visual sound separation using minus-plus net. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 6

[65] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586, 2018. 2, 5, 6

[66] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1735–1744, 2019. 2

[67] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2022. 6

[68] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*, 2023. 6