

CrossMAE: Cross-Modality Masked Autoencoders for Region-Aware Audio-Visual Pre-Training

Yuxin Guo^{1,2,3}, Siyang Sun³, Shuailei Ma^{4,5}, Kecheng Zheng⁴
Xiaoyi Bao^{1,2,3}, Shijie Ma^{1,2}, Wei Zou^{1,2*}, Yun Zheng³

¹School of Artificial Intelligence, University of Chinese Academy of Sciences

²MAIS, Institute of Automation, Chinese Academy of Sciences (CASIA)

³Alibaba Group ⁴Ant Group ⁵Northeastern University

Fguoyuxi n2021, wei . zou6@i a. ac. cn

Abstract

Learning joint and coordinated features across modalities is essential for many audio-visual tasks. Existing pre-training methods primarily focus on global information, neglecting fine-grained features and positions, leading to suboptimal performance in dense prediction tasks. To address this issue, we take a further step towards region-aware audio-visual pre-training and propose CrossMAE, which excels in Cross-modality interaction and region alignment. Specifically, we devise two masked autoencoding (MAE) pretext tasks at both pixel and embedding levels, namely Cross-Conditioned Reconstruction and Cross-Embedding Reconstruction. Taking the visual modality as an example (the same goes for audio), in Cross-Conditioned Reconstruction, the visual modality reconstructs the input image pixels conditioned on audio Attentive Tokens. As for the more challenging Cross-Embedding Reconstruction, unmasked visual tokens reconstruct complete audio features under the guidance of Learnable Queries implying positional information, which effectively enhances the interaction between modalities and exploits fine-grained semantics. Experimental results demonstrate that CrossMAE achieves state-of-the-art performance not only in classification and retrieval, but also in dense prediction tasks. Furthermore, we dive into the mechanism of modal interaction and region alignment of CrossMAE, highlighting the effectiveness of the proposed components.

1. Introduction

Visual and auditory perception are the fundamental abilities of humans to understand the world. Similarly, for intelligent machines, a well-trained audio-visual model can be leveraged in a myriad of applications, such as video pars-

Figure 1. Audio-visual source localization results. It can be observed that existing pre-trained models perform poorly on downstream tasks involving dense audio-visual predictions. However, CrossMAE, due to its ability to concentrate on fine-grained local features of tokens and cross-modal interactions, exhibits outstanding performance in audio-visual source localization tasks.

ing [41, 46], sound source localization [15, 30, 38, 51], navigation [3, 4], sound separation [9, 42, 43], and environment perception [35]. To make audio-visual models strong, various pre-training strategies [11, 12, 14, 17, 20, 44] have been widely used [15, 25, 26, 32, 37, 40, 45].

Specifically, Audio-CLIP [17] and CAV-MAE [14], primarily focus on aligning global features of two modalities, yielding superior results in tasks such as audio-visual classification and retrieval. However, dense prediction audio-visual tasks, such as audio-visual source localization and segmentation, require a nuanced understanding of fine-grained object details and temporal-frequency characteristics of sound, aspects that prior methodologies have largely overlooked. Considering this, a natural question is *whether these models can handle dense prediction tasks like audio-visual source localization?*

To answer this question, we evaluated the performance of several pre-trained models [14, 17] in audio-visual source localization [6, 15, 16, 29, 30, 38, 50]. However, as shown

*Corresponding author.

in Fig. 1, it is evident that they are unable to localize the regions of the sounding objects, which indicates that relying solely on global features for dense prediction tasks yields poor performance. We attribute this gap primarily to the pre-training phase, where models are trained to match the entire image with its paired audio clip without considering the alignment between specific local images and audio regions. So, *how can we incorporate fine-grained information to a pre-trained strategy can reason about regions?*

In this paper, we take a further step towards region-aware audio-visual pre-training and aim to propose an audio-visual pre-trained model that can achieve region alignment and effective cross-modality interaction. Although there is a natural correspondence between the visual images and audio in videos, aligning regions between visual images and audio spectrograms is challenging without annotations. Moreover, how to effectively enable the model to understand and interact with local information at the object level or even token level has provoked our thoughts.

To address these challenges, we propose CrossMAE, which fully explores cross-modal interaction and captures fine-grained region alignment. In this approach, we improve upon masked autoencoders (MAE) [18] to enable effective cross-modal interaction and concentrate on the rich fine-grained positional information encoded by tokens. As described in Fig. 2, we introduce dual levels of MAE tasks in the pixel and embedding space. Firstly, at the pixel level, we enhance the naive MAE by introducing a cross-modal interaction process. Taking the visual modality as an example, unmasked visual tokens are conditioned on Audio Attentive Tokens during the reconstruction of the original input image, namely Cross-Conditioned Reconstruction. Additionally, we propose a more challenging reconstruction task to enhance the model’s perception of region features and modal interaction. At the embedding space level, audio features are directly reconstructed from the visual modality. To achieve region-aware feature learning, we draw inspiration from DETR [2] and introduce Learnable Queries. In DETR, different Learnable Queries represent different objects, while in our approach, different queries represent different positions. During this training phase, audio queries contain positional information of audio features and effectively guide the feature transformation from visual to audio, enabling more effective representation learning of the visual encoder, namely Cross-Embedding Reconstruction. Besides, we also employ contrastive loss to ensure indispensable global alignment.

In CrossMAE, region-aware learning and contrastive training mutually boost each other. Contrastive learning effectively aligns two modalities globally. Attentive Tokens enable the model to concentrate more on aligning regional tokens across modalities, while Cross-Embedding Reconstruction promotes greater attention to positional semantics.

They jointly compensate for the limitation of contrastive learning, which only focuses on the discrimination between positive and negative samples globally but neglects the fine-grained semantics.

In summary, we contribute in the following aspects:

- We propose a versatile audio-visual pre-trained model called **CrossMAE**, which excels in effective **cross-modality** interaction and achieves region-aware representation by employing masked autoencoders (**MAE**).
- A key advantage of our approach lies in the devised dual-level masked autoencoder, including Cross-Conditioned and Cross-Embedding Reconstruction, which facilitates region alignment and modal interaction steered by proposed Attentive Tokens and Learnable Queries.
- CrossMAE achieves state-of-the-art performance in various audio-visual tasks such as classification, retrieval, and dense prediction tasks, which showcases its capability to enhance both single- and cross-modal representation.

2. Related Works

Masked Autoencoders for Single Modality. Vision Transformers [7] treat an image as a sequence of patch tokens and perform encoding computation on these tokens. The masking image modeling strategy [18] randomly masks patches of an image and utilizes the remaining tokens to reconstruct the original pixel values. This approach facilitates the learning of image representations by encouraging the model to comprehend the underlying structure and features of the image through pixel-level recovery. For example, Dynamic ViTs [36] employ token removal to achieve efficient image classification. Similarly, Masked Autoencoder [18] randomly masks 75% of tokens, leading to substantial acceleration and enhancement of self-supervised visual representation pre-training. Taking inspiration from this, MAE-AST [1, 13], AudioMAE [19] introduce the masking strategy to the audio domain by converting audio into spectrogram and treating it as a whole image for reconstruction, achieving state-of-the-art performance on multiple audio classification tasks. In this paper, we extend masked autoencoders to both visual and audio modalities and enhance them with dual-level reconstruction, which endows the model with region alignment capabilities.

Audio-Visual Pre-Training Pre-training plays a crucial role in extracting coordinate features and establishing a foundation for various downstream tasks. However, there is minimal research on audio-visual pre-training. Morgardo proposed employing multiple instances as superior positive sets for contrastive learning and audio-visual representation learning [33]. CAV-MAE [14] was the pioneering work to integrate mask modeling into audio-visual pre-training. By combining contrastive learning with masked autoencoders, it achieved promising results in audio-visual classi-

Figure 2. Overview of the proposed CrossMAE. **(a)**: Overall learning process of CrossMAE, which employs dual-level masked data modeling, including Cross-Conditioned and Cross-Embedding Reconstruction. **(b)**: At the Cross-Conditioned Reconstruction level, we propose a task of reconstructing self-pixels with supervision from the Attentive Tokens of counterpart modality. **(c)**: At the Cross-Embedding Reconstruction level, we propose reconstructing the complete features of the counterpart modality using Learnable Queries with a cross-attention module. We take the visual modality as an example to illustrate (b) and (c), and the same applies to the audio modality.

fication tasks. Subsequently, AV-MAE[11] and MAViL[20] explored the utilization of masked autoencoding to understand the correlation between two modalities, which bears similarities to our approach. We delineate three principal distinctions between CrossMAE and previous methods. **(1)Alignment.** They explore MAE only for global alignment, while we emphasize region alignment. **(2)Methodology.** We explore dual-level cross-modal reconstruction with a separate structure and end-to-end pre-training, differing from the joint masked autoencoding or two-stage training of AV-MAE and MAViL. **(3)Tasks.** Beyond their focus on global tasks like classification and recognition, we additionally realize region or dense prediction tasks. In addition to these approaches, there are other multimodal pre-trained models such as Audio-CLIP [17], Image-Bind [12], and ONE-PEACE [44], which incorporate visual, audio and other modalities. However, these methods primarily focus on global features and overlook the rich spatial, semantic, and frequency information in region features. As a result, they struggle with dense prediction tasks and often require dataset-specific fine-tuning, limiting generalization. In contrast, CrossMAE pays greater attention to region information in both modalities, leading to excellent performance in downstream tasks, especially dense prediction tasks.

Task Specific Audio-Visual Learning. Audio-visual downstream tasks broadly fall into global classification and dense prediction tasks. Global tasks emphasize global fea-

tures like retrieval and classification, while dense prediction tasks utilize local features and positional information from images or audio. For instance, audio-visual source localization [6, 15, 16, 24, 29, 30, 38, 50] involves calculating similarity based on audio and visual features at different positions to generate a final localization map. AVS [28, 32, 51, 52] mainly outputs a pixel-level map of the object(s) producing sound at the time of the image frame. Additionally, there are tasks like AVE [23, 27, 40, 47], AVQA [22, 48, 49], etc. These tasks rely on capturing audio-visual semantics. Moreover, existing methods employ task-specific frameworks, lacking generalization abilities. Considering this, One-AVM [31] proposed a unified model that combines several tasks, achieves unity at the task level and significantly motivates us. In this paper, we propose a region-aware audio-visual pre-trained model that extracts features applicable to both global classification and dense prediction tasks, demonstrating better generalization capabilities.

3. CrossMAE

Overview. In this section, we present the structure and pre-training strategy of CrossMAE, as illustrated in Fig. 2. The pre-training strategy consists of two main components: masked data modeling (MDM) and audio-visual contrastive learning (Sec. 3.1). Specifically, MDM is divided into two

levels of reconstruction. For Cross-Conditioned Reconstruction (Sec. 3.2), unmasked tokens from a single modality reconstruct self-pixels conditioned on Attentive Tokens from the other modality. For Cross-Embedding Reconstruction (Sec. 3.3), the Learnable Queries are employed to guide the unmasked tokens to reconstruct complete features of the other modality. Additionally, CrossMAE employs contrastive loss (Sec. 3.4) to map the features of both modalities into the same space and achieve global alignment. Furthermore, we further explain the meaning and importance of regions of two modalities in Sec. 3.5.

3.1. Preliminary

Videos not only exhibit natural audio-visual correspondence but are also widely accessible in existing network environments. To effectively exploit this natural consistency and facilitate mutual supervision for representation learning between audio and visual modalities, we employ audio-visual contrastive learning and MDM as pre-training strategies. Audio-visual contrastive learning aims to achieve global alignment between two modalities, while MDM emphasizes local features and inter-modality interaction.

Audio-Visual Contrastive Learning Contrastive learning is an effective self-supervised learning approach. It maximizes the similarity between frames and their corresponding audio clips (positive samples) while minimizing the similarity among unpaired ones (negative samples).

$$\begin{aligned} L_{\text{contrast}} = & -E_{(a_i, v_i)} \log \frac{\exp(s(a_i, v_i) / \tau)}{\sum_{j=1}^n \exp(s(a_i, v_j) / \tau)} \\ & + \log \frac{\exp(s(v_i, a_i) / \tau)}{\sum_{j=1}^n \exp(s(v_i, a_j) / \tau)}, \end{aligned} \quad (1)$$

where a and v represent the global features of the audio and visual modalities, respectively. The function $s(\cdot)$ denotes the consistency matching criterion, typically using cosine similarity. Through this approach, the model not only learns effective audio-visual correspondence but also maps the visual and audio features into the same space, thereby reducing the semantic gap between modalities.

Masked Data Modeling (MDM). MDM is another widely used pre-training approach. It divides the input image into multiple patches and randomly masks these patches. The remaining tokens, which are not masked, are then used to reconstruct the original image tokens. Several models have been proposed based on this criterion, and Masked Autoencoders is one of the simple yet highly effective methods among them.

Masked Autoencoders (MAE). MAE primarily utilizes the non-masked tokens to reconstruct the original image

pixels [18], thereby enhancing the correlation between image patches and aiding the model's spatial awareness.

$$L_{\text{recon}}(\hat{v}, v) = \frac{1}{N} \|\hat{v} - v\|_2^2, \quad (2)$$

where \hat{v} is the reconstructed patches, and v is the original pixel patches. Currently, MAE is widely employed in various unimodal and multimodal representation learning tasks, achieving impressive results. In this paper, we extend the naive MAE to a dual-level reconstruction, enabling mutual supervision between two modalities to strengthen the correlation between them.

3.2. Cross-Conditioned Reconstruction

To ensure self-representation capabilities in both modalities, we allow them to reconstruct their modal pixels. To enhance the consistency between modalities, we improve the single-modal MAE by introducing Cross-Conditioned Reconstruction, which allows each modality to reconstruct its own pixels conditioned on the tokens with the highest attention from the other modality.

Random Masking. Given a batch of N unlabeled audio-visual pairs, we first transform the raw audio waves into spectrograms as the model inputs $\{(A_i, V_i), i = [1, N]\}$. Before entering the Encoder, we tokenize the frame V and spectrogram A separately, add two-dimensional position embedding, and randomly mask them with a high mask ratio (always 75%).

$$A_i[\text{vis}] = \text{MASK}(\text{Patch.Emb}(A_i) + E_{\text{pos}}^a), \quad (3)$$

$$V_i[\text{vis}] = \text{MASK}(\text{Patch.Emb}(V_i) + E_{\text{pos}}^v), \quad (4)$$

Where $A_i[\text{vis}]$ and $V_i[\text{vis}]$ represent the unmasked tokens in the frame and spectrogram of the i -th sample, respectively. $\text{MASK}(\cdot)$ denotes the random masking operation applied to the patches. Subsequently, we input the unmasked patches $A_i[\text{vis}]$, $V_i[\text{vis}]$ into the encoder and obtain their features $a_i[\text{vis}]$ and $v_i[\text{vis}]$.

Attentive Tokens. Here, we select the tokens with the highest attention to the $[\text{CLS}]$ token in each modality as Attentive Tokens, which will guide the pixel reconstruction of the other modality. We output the attention of different tokens in the last layer block of the encoder and sort them. Then, we select the top 25% of tokens as Attentive Tokens $a_i[\text{attn}]$, $v_i[\text{attn}]$:

$$a_i[\text{attn}] = \text{topK}(\text{sim}([\text{CLS}]_i^a, a_i[\text{m}]), k), \quad (5)$$

$$v_i[\text{attn}] = \text{topK}(\text{sim}([\text{CLS}]_i^v, v_i[\text{n}]), k), \quad (6)$$

where a_i, v_i are feature tokens obtained from encoders, and m and n represent the number of visual and audio patches, respectively. $\text{topK}(x, k)$ refers to selecting the top k tokens from x , where k is equal to 25% of the patch number.

Conditioned Reconstruction. Finally, we allow each modality to reconstruct its image (spectrogram) pixels conditioned on the Attentive Tokens of the other modalities. We introduce a cross-attention layer into the decoder to encode the Attentive Tokens into features themselves:

$$a_i[\text{cdt}] = \text{CrossAttn}(a_i[\text{vis}], v_i[\text{attn}], v_i[\text{attn}]), \quad (7)$$

$$v_i[\text{cdt}] = \text{CrossAttn}(v_i[\text{vis}], a_i[\text{attn}], a_i[\text{attn}]), \quad (8)$$

where $a_i[\text{cdt}]$, $v_i[\text{cdt}]$ represents token features that have been updated by Attentive Tokens from the counterpart modality. Afterwards, we pass the $a_i[\text{cdt}]$, $v_i[\text{cdt}]$ through the decoder to obtain reconstructed pixel \hat{A}_i , \hat{V}_i . Then, we calculate the reconstruction loss between \hat{A}_i , \hat{V}_i , and the unmasked original pixels A_i , V_i using formula 2.

$$L_{\text{cross-cdt}} = L_{\text{recon}}(\hat{A}_i[\text{vis}], A_i[\text{vis}]) + L_{\text{recon}}(\hat{V}_i[\text{vis}], V_i[\text{vis}]), \quad (9)$$

where $[\text{vis}]$ represents only calculating the reconstruction loss for the image patches that have not been masked. Through Cross-Conditioned Reconstruction, the model not only improves the representation capability of each modality individually but also complements the contrastive loss, strengthening the region modeling between the two modalities, which can be demonstrated in Sec. 4.3.

3.3. Cross-Embedding Reconstruction

To further enhance the model's cross-modal region modeling capability, we propose a more challenging task: Cross-embedding reconstruction. This task requires the non-masked tokens, guided by Learnable Queries, to generate complete features of the other modality. We hope that this task can serve as a supportive role for the Cross-Conditioned Reconstruction task. Taking the visual modality as an example in this task, we set it at the embedding level. By setting Learnable Queries with the same dimensionality as the audio modality that needs to be restored, audio queries are enabled to imply positional information of audio features, guiding the visual modality's unmasked tokens in feature transformation effectively. The same approach applies to the audio modality.

Learnable Queries. For the visual modality, we draw inspiration from the principles of DETR [2]. In DETR, different queries represent different objects. In this case, different Learnable Queries represent various positions, aiming to leverage structural information from the audio features to facilitate the conversion of visual unmasked tokens into complete audio features. We set the visual and audio Learnable Queries as q_v and q_a , respectively, where the dimension of q_v is the same as the visual feature dimension, and the same applies to q_a .

Cross-Embedding Decoder. We develop a Cross-Embedding Reconstruction decoder that combines the

Learnable Queries with the modality-specific features to generate complete features of the other modality. Structurally, each decoder consists of several blocks, which contain the self-attention and cross-attention layer to combine the Learnable Queries and features, and a feed-forward network (FFN) layer. The final cross-modal reconstructed features are obtained as follows:

$$\hat{a}_i = \text{Dec}(\text{CrossAttn}(q_a, v_i[\text{vis}], v_i[\text{vis}])), \quad (10)$$

$$\hat{v}_i = \text{Dec}(\text{CrossAttn}(q_v, a_i[\text{vis}], a_i[\text{vis}])), \quad (11)$$

where $\text{Dec}(\cdot)$ represents decoder, $\text{CrossAttn}(Q, K, V)$ is the cross-attention layer in the decoder. In each block of the decoder, there is a self-attention layer (which is not explicitly represented in the formula), a cross-attention layer, and a feedforward layer. Finally, we minimize the mean square error between the reconstructed features and the normalized complete original features.

$$L_{\text{cross-emb}} = L_{\text{recon}}(\hat{a}_i, a_i) + L_{\text{recon}}(\hat{v}_i, v_i). \quad (12)$$

By completing this challenging cross-modal feature generation task, the model not only generates features for the other modality but also gains a more comprehensive understanding of the positional and semantic information of its own modality. This, in turn, promotes the learning of joint and coordinated features in our pre-trained model.

3.4. Overall Learning Objectives

In addition to aligning region features, we also employ audio-visual contrastive learning loss L_{contrst} , as shown in Eq. 1, to ensure the fundamental representation capabilities and global alignment of both modalities. Therefore, we can finally get the overall learning objectives L_{CrossMAE} as:

$$L_{\text{CrossMAE}} = L_{\text{contrst}} + L_{\text{cross-cdt}} + L_{\text{cross-emb}}. \quad (13)$$

3.5. Further Explanation of Regions

We need to clarify the significance of region features for both modalities. In the visual modality, regions correspond to the positions and semantics of different objects in frames, enabling tasks like localization. In the case of the audio modality, audio is typically transformed into spectrograms, where the horizontal axis represents time and the vertical axis represents frequency. From a signal perspective, the spectrogram reflects the energy distribution and temporal changes in different frequency bands of the audio signal, while regions indicate specific information within certain time or frequency segments, providing a certain correspondence to sound sources or events. Therefore, region features are crucial for the perception of fine-grained visual and audio characteristics.

Table 1. Performance of audio-visual classification on three fine-tuned datasets. For the AudioSet dataset, we report mAP while accuracy for VGGSound. SL = supervised learning; SSL = self-supervised learning; IN = ImageNet; IN21K = ImageNet-21K; AS = AudioSet-2M; VGG = VGGSound. $\bar{\cdot}$ denotes non-standard pre-trained datasets, like ImageNet-21K (much larger than ImageNet) or VGGSound.

Methods	Pre-training	AudioSet-20K			AudioSet-2M			VGGSound-200K		
		A	V	A-V	A	V	A-V	A	V	A-V
(a) Audio-CLIP [17]	-	25.4	12.8	35.4	34.1	25.2	41.2	49.7	44.2	55.8
(b) Perceiver [21]	-	-	-	-	38.4	25.7	44.2	-	-	-
(c) Attn AV [8]	SL + IN	-	-	-	38.4	25.7	46.2	-	-	-
(d) MBT [34]	SL + IN21K	31.3	27.7	43.9	44.3	32.3	52.1	52.3	51.2	64.1
(e) CAV-MAE [14]	SSL + IN, AS	<u>37.7</u>	<u>19.8</u>	42.0	<u>46.6</u>	26.2	51.2	<u>59.5</u>	<u>47.0</u>	<u>65.5</u>
(f) AV-MAE [11]	SSL + IN, AS/VGG	-	-	-	<u>46.6</u>	31.1	51.8	<u>57.2</u>	50.3	<u>65.0</u>
(g) CL-AV	SSL + IN, AS	31.9	16.6	38.2	43.8	23.8	46.1	54.7	45.3	57.1
(h) SelfMAE	SSL + IN, AS	29.8	14.9	36.5	42.5	23.6	39.6	47.5	43.6	49.8
(i) CL-SelfMAE	SSL + IN, AS	32.6	17.1	41.3	44.2	24.4	48.2	55.9	46.1	61.7
(j) CL-CrossCdtMAE	SSL + IN, AS	33.6	17.8	44.9	45.3	25.3	51.9	57.1	46.9	62.9
(k) CL-CrossEmbMAE	SSL + IN, AS	36.9	19.0	<u>47.7</u>	46.4	26.5	<u>54.4</u>	59.3	<u>47.5</u>	64.4
(l) CrossMAE (ours)	SSL + IN, AS	39.2	20.4	48.2	47.1	<u>27.2</u>	55.3	61.1	48.2	67.0

4. Experiments

With the dual-level reconstruction, under the same experimental setup (Sec. 4.1), CrossMAE not only achieves excellent single- and cross-modal representations, but also demonstrates strong region alignment (Sec. 4.3) compared to different baseline variants (Sec. 4.2). Especially, we focus our attention on the following issues in our Ablation and Further Analysis in Sec. 4.4 and validate them:

- What about the scalability of CrossMAE?
- How do global and region alignment promote each other?
- What is the key to the effectiveness of Attentive Tokens?
- Why are the Learnable Queries quite indispensable?

4.1. Experimental Setup

Datasets. We utilized three datasets: AudioSet [10], VGGSound [5], and Flickr-SoundNet [38, 39]. AudioSet [10] consists of YouTube videos that cover 527 different sound events. VGGSound [5] comprises 200K 10-second YouTube video clips annotated with 309 classes. FlickrSoundNet [38, 39] and VGG-SoundSource [6] contains 5,000 and 5,158 bounding-box annotations, respectively. We pre-trained CrossMAE on AudioSet-2M. For event classification, we fine-tuned models on AudioSet and VGGSound-200K. For retrieval and AVSL, we utilized Flickr-SoundNet and VGG-SoundSource for the evaluation. More details are in the Appendix.

Implementation Details. For the visual and audio architectures, we follow the settings of MAE [18] and AudioMAE [19], respectively. Our framework utilizes a ViT-B/16 model [7] as the backbone by default. Models are pre-trained for 30 epochs with a batch size of 512. For input, the image is 224x224, and the spectrogram is 1024x128. Patch size is 16x16. We set $\beta = 0.1$ to scale the gradients from

Table 2. Different variants (baselines) of CrossMAE.

Variants	contrastive loss	vanilla recon loss	visual recon loss	vanilla audio recon loss	cross-condition recon loss	cross-embedding recon loss
CL-AV						
SelfMAE						
CL-SelfMAE						
CL-CrossCdtMAE						
CL-CrossEmbMAE						
CrossMAE						

losses into a comparable range to improve training stability. Moreover, following MAE, we set 0.75 for random masking ratio, and 0.25 for Attentive Tokens’ top ratio.

4.2. Baseline Variants of CrossMAE

To validate the effectiveness of CorssMAE, we consider the following baseline variants as below (Table 2).

- **CL-AV:** Only audio-visual contrastive learning.
- **SelfMAE:** Self-modality MAE reconstruction.
- **CL-SelfMAE:** Audio-visual contrastive learning and self-modality MAE reconstruction.
- **CL-CrossCdtMAE:** Audio-visual contrastive learning and Cross-Conditioned Reconstruction.
- **CL-CrossEmbMAE:** Audio-visual contrastive learning and Cross-Embedding Reconstruction.
- **CrossMAE:** Full version as described in Section 3.

4.3. Comparison With State-of-the-Art Methods

Significant improvement in single- and cross-modal representations. We evaluate the performance of CrossMAE on single- and multi-modal classification and compare it with previous methods. We fine-tuned the model with randomly initialized linear classification heads. Specifically, we evaluate the performance of audio-visual event classification tasks using three different datasets: (1) with similar distribution but different samples (AudioSet-20K), (2) with the same distribution and identical samples (AudioSet-2M),

Table 3. Performance of audio-visual source localization on Flickr-SoundNet and VGG-SoundSource.

Methods	Flickr		VGG-SoundSource	
	CIoU	AUC	CIoU	AUC
(a) CAV-MAE [14]	50.80	30.83	25.20	19.09
(b) CL-AV	65.60	32.08	30.80	22.36
(c) SelfMAE	33.40	21.30	17.77	13.03
(d) CL-SelfMAE	67.20	38.93	31.08	22.28
(e) CL-CrossCdtMAE	68.53	39.72	37.60	25.45
(f) CL-CrossEmbMAE	69.11	40.62	38.20	25.94
(g) CrossMAE (ours)	73.20	42.61	39.80	27.13

Table 4. Scaling experiments on different model/batch sizes.

Backbones	batch size	Audio-Visual			Visual-Audio			AVSL	
		R@1	R@5	R@10	R@1	R@5	R@10	CIoU	AUC
ViT-B/16	256	20.4	52.8	63.4	18.6	51.8	57.2	38.06	25.62
ViT-B/16	512	22.8	56.6	70.8	21.3	54.4	60.5	39.80	27.13
ViT-L/16	256	36.4	69.2	78.4	32.8	67.6	77.2	41.68	29.60
ViT-L/16	512	37.2	70.8	80.0	30.4	65.2	79.4	43.04	30.92

and (3) with a different distribution (VGGSound-200K). We fine-tuned the model using audio-only, video-only, and audio-visual data to evaluate its performance in both single-modal and multi-modal scenarios. The experimental results are shown in the Table 1. Compared to prior methods (a-f), CrossMAE (l) significantly improves both single-modal and cross-modal representations. Notably, against naive global interactions from contrastive learning (g, i), our proposed region interactions like Cross-Conditioned and Cross-Embedding Reconstruction (j-l) remarkably enhance model representation capability, highlighting the effectiveness of our method.

Stronger regional-aware representations in dense prediction tasks. We evaluated various pre-training methods on audio-visual source localization and reported Consensus Intersection over Union (CIoU) and Area Under Curve (AUC) metrics. In Fig. 1 and Table 3, prior pre-training methods struggle to accurately localize sounding objects. In contrast, CrossMAE improves AUC by 11.78% on Flickr. Both Attentive Tokens and Cross-Embedding Reconstruction contribute significant improvements to CrossMAE. We attribute this to the fact that both components guide the model’s attention to valuable tokens and enable the capture of fine-grained positions.

4.4. Ablation and Further Analysis

What about the scalability of CrossMAE? We conducted scaling experiments with different model sizes and batch sizes. Results in Table 4 show that the performance gains of CrossMAE increase with larger batch size, and are more obvious when the backbone is expanded from ViT-B/16 to ViT-L/16, which shows that CrossMAE is a scalable pre-training scheme.

Table 5. Performance of audio-visual retrieval results on Flickr-SoundNet. We report R@1, R@5, and R@10 results.

	Audio		Visual		Visual		Audio	
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@10
(a) CAV-MAE [14]	12.2	26.0	38.2	18.2	28.8	42.5		
(b) CL-AV	10.4	24.4	40.4	7.6	22.8	38.4		
(c) SelfMAE	2.0	3.6	12.8	1.8	3.2	12.0		
(d) CL-SelfMAE	11.8	26.0	42.8	8.4	24.0	38.0		
(e) CL-CrossCdtMAE	13.6	34.8	46.4	13.2	36.0	50.8		
(f) CL-CrossEmbMAE	14.2	39.6	53.4	13.6	32.8	52.8		
(g) CrossMAE (ours)	16.8	45.2	64.0	18.2	46.4	60.8		

How do global and region alignment promote each other? (1) We first analyze the impact of global alignment by studying the effect of contrastive learning. Table 3 of audio-visual source localization reveals that solely using self-modality MAE (g) is weaker than (f,h) about 6.5% and 8.6%. The results highlight that contrastive learning unifies two modalities within a shared feature space, which is fundamental for further local alignment and feature reconstruction. (2) Regarding the influence of local alignment, we conducted experiments on audio-visual retrieval without fine-tuning. Table 5 demonstrates that region alignment from Attentive Tokens and Cross-Embedding Reconstruction (e-g) significantly enhances global alignment (b, d). We attribute these performance improvements to the fact Attentive Tokens enable the model to concentrate more on aligning regional tokens across modalities, while Cross-Embedding Reconstruction promotes greater attention to positional and semantics. They jointly compensate for the limited global alignment of contrastive learning.

What is the key to the effectiveness of Attentive Tokens?

To delve deeper into the Attentive Tokens, we conducted ablations on the formulation of Attentive Tokens. We replaced the Attentive Tokens with empty vectors and random tokens and evaluated their performance in retrieval and audio-visual source localization tasks. As shown in Table 6, empty tokens did not improve performance, while random tokens limited performance owing to containing much irrelevant information. Attentive tokens, as the most attended part of the [CLS] tokens, carrying rich positional semantics, could largely enhance modality interaction.

For more clarity, we visualize visual Attentive Tokens of some images (the same goes for audio) and find they primarily highlight the sounding objects, depicted as red tokens in Fig. 3. These tokens help the reconstruction of audio spectrograms under the guidance of fine-grained visual information of the sounding objects, integrating local visual details into audio and thereby achieving effective region alignment.

Why are the Learnable Queries quite indispensable?

To analyze the key to the success of Cross-Embedding Reconstruction, we replaced Learnable Queries with various

Figure 3. Qualitative analysis: visualization results of Attentive Tokens and Learnable Queries. (a) Red tokens represent Attentive Tokens, which predominantly cover sounding objects. (b) Audio Learnable Queries, where different queries focus on various positions of audio features. (c) Visual Learnable Queries, where different queries focus on different salient regions in the image.

Table 6. Detailed ablations of Attentive Tokens on several tasks.

Methods	Audio		Visual		Visual		Audio		AVSL	
	R@1	R@5	R@10	R@1	R@5	R@10	CloU	AUC		
CL-CrossEmbMAE	19.4	48.6	61.2	18.8	47.2	59.6	38.22	25.94		
+ zero vectors	19.2	48.8	60.9	18.6	46.8	59.2	38.24	25.92		
+ random tokens	19.6	49.6	62.6	19.2	50.8	63.4	38.96	26.27		
+ Attentive Tokens	22.8	56.6	70.8	21.3	58.4	67.8	39.80	27.13		

counterparts. Table 7 reveals that using tokens from the other modality no longer works in this case, because these tokens introduce a large amount of information about the reconstruction target, which significantly reduces the difficulty of the pretext task and makes it ineffective. However, when replacing the tokens with queries containing less information helps alleviate this issue. Unlike empty queries, Learnable Queries maintain useful information and help feature transformation.

Furthermore, we delved deeper into what information Learnable Queries acquire and what role they play in Fig. 3. We visualized each query and summarized three key points: **(1). Queries guide the generation of audio features at specific locations.** As illustrated in Fig. 3(b), we visualize the attention maps of queries and the reconstructed audio features, finding that different queries focus on various specific locations, which do not change with sample variations. **(2). Queries uncover region-aligned information between two modalities.** Fig. 3(c) displays attention maps of queries and original visual features, showing different queries focus on various principal regions of the image. Combined with (1), we observe that queries mainly focus on similar positions in the spectrogram, and tend to target the same subject objects in the image. **(3). Queries pre-define and guide the dimension of cross-modal feature reconstruction,** addressing the issue of differing quantities of visual and audio tokens.

Table 7. Detailed ablations of Learnable Queries on several tasks.

Methods	Audio		Visual		Visual		Audio		AVSL	
	R@1	R@5	R@10	R@1	R@5	R@10	CloU	AUC		
CL-CrossCdtMAE	15.2	42.4	52.8	14.0	41.2	51.6	37.60	25.45		
+ Attentive Tokens	15.8	43.6	54.6	15.4	42.8	52.4	37.80	25.83		
+ empty query	17.6	45.2	60.0	17.2	44.0	59.2	38.20	26.60		
+ Learnable Queries	22.8	56.6	70.8	21.3	58.4	67.8	39.80	27.13		

5. Conclusion

In this paper, we propose a region-aware universal audio-visual pre-trained model, namely CrossMAE, which possesses excellent modality interaction and region alignment capabilities. We introduce two proxy tasks, Cross-Conditioned Reconstruction and Cross-Embedding Reconstruction, to enhance the interaction between the two modalities and effectively capture fine-grained positional and semantic information. As a result, our model achieves outstanding performance on classification tasks such as audio-visual classification and audio-visual retrieval, particularly excelling in dense prediction tasks like audio-visual source localization. Moreover, we investigate the improvement in audio-visual representation learning through modality interaction and region alignment, demonstrating the effectiveness and training efficiency of our model.

We hope this work will help draw more attention to audio-visual pre-training, and provoke a reconsideration of modality interaction and region-perception, to stimulate more research in this challenging yet significant task.

Acknowledgements. This work has been supported by the National Natural Science Foundation of China under Grant 61773374 and the Major Basic Research Projects of Natural Science Foundation of Shandong Province under Grant ZR2019ZD07.

References

- [1] Alan Baade, Puyuan Peng, and David Harwath. Mae-ast: Masked autoencoding audio spectrogram transformer. *arXiv preprint arXiv:2203.16691*, 2022. **2**
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. **2, 5**
- [3] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Venc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 17–36. Springer, 2020. **1**
- [4] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15516–15525, 2021. **1**
- [5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggssound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. **6**
- [6] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. **1, 3, 6**
- [7] 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 *International Conference on Learning Representations*, 2021. **2, 6**
- [8] Haytham M Fayek and Anurag Kumar. Large scale audio-visual learning of sounds with weakly labeled data. *arXiv preprint arXiv:2006.01595*, 2020. **6**
- [9] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15490–15500. IEEE, 2021. **1**
- [10] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. **6**
- [11] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16144–16154, 2023. **1, 3, 6**
- [12] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. **1, 3**
- [13] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. **2**
- [14] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2023. **1, 2, 6, 7**
- [15] Yuxin Guo, Shijie Ma, Hu Su, Zhiqing Wang, Yuhao Zhao, Wei Zou, Siyang Sun, and Yun Zheng. Dual mean-teacher: An unbiased semi-supervised framework for audio-visual source localization. *Advances in Neural Information Processing Systems*, 36, 2024. **1, 3**
- [16] Yuxin Guo, Shijie Ma, Yuhao Zhao, Hu Su, and Wei Zou. Cross pseudo-labeling for semi-supervised audio-visual source localization. *arXiv preprint arXiv:2403.03095*, 2024. **1, 3**
- [17] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. **1, 3, 6**
- [18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. **2, 4, 6**
- [19] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35: 28708–28720, 2022. **2, 6**
- [20] Po-Yao Huang, Vasu Sharma, Hu Xu, Chaitanya Ryali, Yanghao Li, Shang-Wen Li, Gargi Ghosh, Jitendra Malik, Christoph Feichtenhofer, et al. Mavil: Masked audio-video learners. *Advances in Neural Information Processing Systems*, 36, 2024. **1, 3**
- [21] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. **6**
- [22] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. **3**
- [23] 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*, 2020. **3**
- [24] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *Proceedings of the 30th*

- ACM International Conference on Multimedia*, pages 3742–3753, 2022. **3**
- [25] Shijie Ma, Fei Zhu, Zhen Cheng, and Xu-Yao Zhang. Towards trustworthy dataset distillation. *arXiv preprint arXiv:2307.09165*, 2023. **1**
- [26] Shijie Ma, Fei Zhu, Zhun Zhong, Xu-Yao Zhang, and Cheng-Lin Liu. Active generalized category discovery. *arXiv preprint arXiv:2403.04272*, 2024. **1**
- [27] Tanvir Mahmud and Diana Marculescu. Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5158–5167, 2023. **3**
- [28] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*, 2023. **3**
- [29] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022. **1, 3**
- [30] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Advances in Neural Information Processing Systems*, 2022. **1, 3**
- [31] Shentong Mo and Pedro Morgado. A unified audio-visual learning framework for localization, separation, and recognition. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. **3**
- [32] Shentong Mo and Bhiksha Raj. Weakly-supervised audio-visual segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. **1, 3**
- [33] 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*, pages 12475–12486, 2021. **2**
- [34] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. **6**
- [35] Janani Ramaswamy. What makes the sound?: A dual-modality interacting network for audio-visual event localization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4372–4376, 2020. **1**
- [36] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Advances in Neural Information Processing Systems*, 2021. **2**
- [37] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *CVPR*, 2023. **1**
- [38] 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*, pages 4358–4366, 2018. **1, 3, 6**
- [39] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound sources in visual scenes: Analysis and applications. *TPAMI*, 43(5): 1605–1619, 2019. **6**
- [40] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. **1, 3**
- [41] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multi-sensory perception: Weakly-supervised audio-visual video parsing. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 436–454. Springer, 2020. **1**
- [42] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. *arXiv preprint arXiv:2011.01143*, 2020. **1**
- [43] Efthymios Tzinis, Scott Wisdom, Tal Remez, and John R Hershey. Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation. In *European Conference on Computer Vision*, pages 368–385. Springer, 2022. **1**
- [44] 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 arXiv:2305.11172*, 2023. **1, 3**
- [45] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022. **1**
- [46] Yu Wu and Yi Yang. Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1326–1335, 2021. **1**
- [47] Yu Wu, Linchao Zhu, Yan Yan, and Yi Yang. Dual attention matching for audio-visual event localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6292–6300, 2019. **3**
- [48] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*, pages 3480–3491, 2022. **3**
- [49] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2031–2041, 2021. **3**
- [50] 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. **1, 3**
- [51] 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 *European Conference on Computer Vision*, pages 386–403. Springer, 2022. [1](#), [3](#)

- [52] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*, 2023. [3](#)