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

Supplementary Material

Contents

The appendix consists of **baselines, experimental settings and supplementary experiments**. The main contents are listed as follows:

- Section A. **Experimental Details:** we introduce the comparative methods more comprehensively, discuss the similarity and relationship with AV-MAE and describe the implementation details and metrics for audio-visual source localization.
- Section B. **Impact of Top Ratio of Attentive Tokens:** we provide experimental results with various ratios of attentive tokens for both modalities.
- Section C. **Impact of Mask Strategy and Mask Ratio:** more detailed results of the mask mechanism.
- Section D. **More Audio-Visual Retrieval Results:** including experiments on VGGSound and MSR-VTT.
- Section E. **Comparison with Task-Specific Methods of Audio-Visual Source Localization:** we compare the pre-trained models using CrossMAE with task-specific methods in audio-visual source localization.
- Section F. **Comparison with MAViL:** we compare our method with more recent work MAViL [**?**].
- Section G. **Further Audio-Visual Tasks:** we fine-tune the pre-trained model on AVE, AVQA, and AVS tasks.
- Section H. **Limitations:** we enumerate some limitations of our approach.

A. Experimental Details

A.1. Comparative Methods

In the main paper, we compared various existing methodologies, including AudioCLIP, Perceiver, AV Fusion, MBT and CAV-MAE.

- **AudioCLIP** [8] (ICASSP_{2022}): This method extends the CLIP model, presenting an extension of the contrastive text-image model (CLIP) that accommodates audio in addition to text and images. AudioCLIP is an approach that integrates a high-performance audio model into CLIP, thereby achieving a tri-modal hybrid architecture. It is noteworthy that AudioCLIP is a visual-text-audio tri-modal pre-trained model. In our tests, we utilized only the visual and auditory modalities without adjusting the linguistic modality.
- **Perceiver** [10] (ICML_{2021}): The paper introduces a model that utilizes an asymmetrical attention mechanism to progressively refine inputs into a condensed latent bottleneck, enabling the handling of substantially large inputs. It implements a technique for marking input units with a precise representation of position and modality, akin to the labeled line strategy employed in the creation of topographic and cross-sensory maps within biological neural networks, which correlates the activity of a distinct unit with a semantic or spatial location.
- **Attn AV** [3] (IJCAI_{2021}): The paper discusses an audiovisual fusion model designed to integrate separately trained individual audio and visual models for each respective modality. The objective is to identify sounds from weakly labeled video recordings through the use of an attention module. The models are trained using data that is weakly labeled, which enables the system to learn from examples that do not have precise annotations or timestamps, focusing instead on the overall context to make associations between audio and visual elements.
- **MBT** [15] (NeurIPS_{2021}): This research presents a new transformer-based structure incorporating 'fusion bottlenecks' for merging modalities at various levels within the network. Unlike the standard pairwise self-attention, the Multimodal Bottleneck Transformer (MBT) constrains the flow of information between different modalities to pass through a limited set of bottleneck latent. This design compels the model to summarize and distill the pertinent information within each modality and exchange only what is essential. **It is important to highlight that, whereas MBT employs labels for supervised training, our approach adopts self-supervised training, forgoing the need for any annotations.** This allows our model to learn from the data itself, potentially making it more versatile in scenarios where labeled data is scarce or unavailable.
- **CAV-MAE** [5] (ICLR_{2023}): This paper pioneers the expansion of the MAE (Masked Autoencoder) model from its original single-modality application to encompass audio-visual multi-modalities. It puts forward CAV-MAE, an innovative approach that merges contrastive learning with masked data modeling to derive a cohesive audio-visual representation. It operates with a joint encoder and decoder that handles separate audio and visual features as well as a combined feature set derived from concatenating the two. Our proposed method, in contrast, treats audio and visual inputs independently and has been crafted to include a two-tiered reconstruction process. This allows for a more granular and precise reconstruction of each modality, which could potentially lead to a more robust and accurate multi-modal representation.

A.2. Similarity and Distinctions of AV-MAE

Our Cross-Conditioned Reconstruction *v.s.* AV-MAE. Similarity: Both aim to reconstruct one modality using features from the other.

Distinctions: **(1)Pre-training**. We employ advanced interaction, *i.e.*, multi-head cross-attention, and capitalize on regionally correlated information (Attentive Tokens) for reconstruction, enabling fine-grained region alignment. Conversely, AV-MAE uses a basic concatenation of tokens and disregards regional specifics. **(2)Backbone**. We use separate encoders and decoders for each modality, providing more flexible modeling.

A.3. Implementation Details

Dataset Details. AudioSet [4] is the largest dataset for sound events. It consists of YouTube videos that cover 527 different sound events. Each video clip is approximately 10 seconds long and has been manually annotated by humans with multiple labels indicating the sound events present in the clip. Flickr-SoundNet [16, 17] and VGG-SoundSource [2] are two large-scale audio-visual datasets. Flickr-SoundNet contains 5,000 bounding-box annotations, and VGG-SoundSource has 5,158 annotated samples. For audio event classification, we fine-tuned the models using AudioSet-20k, AudioSet-2M, and VGGSound-200K [1] datasets. Additionally, we utilize Flickr-SoundNet and VGG-SoundSource for the evaluation of audio-visual retrieval and audio-visual source localization tasks [2, 6, 7, 13, 14, 16, 23].

Our backbone architecture follows that of the Vision Transformer (ViT). During the training process, we utilize the Adam optimizer with a learning rate set to 2e-5 and a weight decay of 0.05. The training workload is distributed across 8 Nvidia A100 GPUs.

In our audio-visual classification experiments, we fine-tune the pre-trained model on different datasets (for details, see Section 4.3 in the main paper). The remaining tasks, which include audio-visual retrieval and audio-visual source localization, are directly tested without any further fine-tuning. This approach demonstrates the versatility and transferability of the learned representations from our model, indicating its potential effectiveness in a variety of audio-visual tasks even without task-specific fine-tuning.

For images, we extract the middle frame from a 10-second Audioset video. The image is resized to 224×224×3 for input and tokenized with a patch size of 16. We apply a 2D sin-cos position encoding to different tokens. Following MAE, we set the default mask ratio to 0.75 using random masking. For Attentive Tokens, we set the top ratio to 0.25.

For audio, following AST and Attention Bottlenecks, we transform the raw waveform (pre-processed as a mono channel with a 16,000 Hz sampling rate) into 128 Mel-frequency bands compatible with Kaldi, using a 25ms Hanning window with a 10ms shift. We use a 10-second recording from AudioSet and set the input as a spectrogram with 1×1024×128 dimensions. The spectrogram's mask ratio is set to 0.75, and we conduct ablation experiments on different masking methods in this appendix. Similar to vision, the top ratio for audio-Attentive Tokens is 0.25.

Table 1. Performance of various tasks across different visual top ratios, we fix the audio top ratio as 25%. We conduct audio-visual retrieval tests on the VGG-SoundSource dataset and report the results.

V ratio	audio-visual			visual-audio			AVSL	
	R@1	R@5	R@10	R@1	R@5	R@10	CIoU	AUC
0%	19.2	48.8	61.2	18.8	47.2	59.6	38.22	25.94
25%	22.8	56.6	70.8	21.3	54.4	60.5	**39.80**	**27.13**
50%	**24.0**	**58.4**	**72.2**	**22.6**	**56.6**	**62.0**	35.86	25.63
75%	21.2	55.2	68.2	20.6	53.4	59.8	35.46	25.15

Table 2. Performance of various tasks across different audio top ratios, we fix the visual top ratio as 25%.

A ratio	audio-visual			visual-audio			AVSL	
	R@1	R@5	R@10	R@1	R@5	R@10	CIoU	AUC
0%	19.2	48.8	61.2	18.8	47.2	59.6	38.22	25.94
25%	22.8	56.6	**70.8**	21.3	54.4	**60.5**	**39.80**	**27.13**
50%	**23.0**	**57.0**	68.6	**54.6**	**56.0**	60.0	35.85	25.28
75%	21.0	55.2	66.4	19.2	54.2	57.2	34.26	24.65

A.4. Metrics of Audio-Visual Source Localization

We report the Area Under Curve (AUC) and Consensus Intersection over Union (CIoU), following previous settings. We consider a set of audio-visual pairs as $\mathcal{D} = \{(v_i, a_i), \mathcal{G}_i\}$, where $\mathcal{G}_i$ is the ground-truth. We set $\mathcal{P}_i(\delta) = \{(x, y)|\mathcal{P}_i(x, y) > \delta\}$ is the foreground region of predicted map, and $\mathcal{G}_i(x, y) = \{(x, y)|\mathcal{G}_i(x, y) > 0\}$ is the foreground region of ground truth.

The IoU of the predicted map and ground truth can be calculated by:

$$IoU_i(\delta) = \frac{\sum_{x,y\in\mathcal{P}_i(\delta)} \mathcal{G}_i(x, y)}{\sum_{x,y\in\mathcal{P}_i(\delta)} \mathcal{G}_i(x, y) + \sum_{x,y\in\{\mathcal{P}_i(\delta)-\mathcal{G}_i\}} 1}. \tag{1}$$

In previous works, CIoU quantifies the proportion of samples with IoU values exceeding a predetermined threshold.

B. Impact of Top Ratio of Attentive Tokens

In the main text, we explore the influence of attentive token forms on model performance, proving their effectiveness. Additionally, we examine the performance of Attentive Tokens at different top ratios. Specifically, we compare the impact of selecting the top 25%, 50%, and 75% of tokens from both modalities on model performance. We report their performance on the zero-shot audio-visual retrieval task. Results in Table 1 and Table 2 show that a top

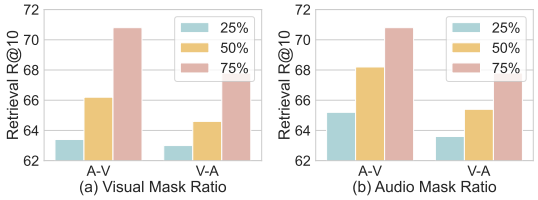

Figure 1. Mask ratios ablation studies. We conducted audio-visual retrieval tests on the VGG-SoundSource dataset and reported the R@10. A mask ratio of 75% is suitable for both visual and auditory modalities in CrossMAE. Therefore, we set the visual and auditory mask ratios of CrossMAE to 75%.

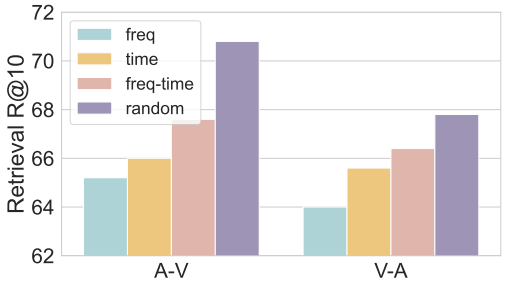

Figure 2. Mask strategies ablation studies. We conducted audio-visual retrieval tests on the VGG-SoundSource dataset and reported the R@10. We found that random masking (unstructured) yields the best results for spectrograms, followed by time-freq masking (structured), while masking only time or frequency individually leads to slightly lower performance. We believe that the random and time-freq masking methods are more conducive to the model's reconstruction based on information from surrounding tokens.

ratio of 25% yields the best audio-visual source localization effects for CrossMAE and sufficiently good audio-visual retrieval results, surpassing existing state-of-the-art method CAV-MAE. At a top ratio of 50%, there is a slight improvement in audio-visual retrieval performance; however, it is minimal and the R@10 metrics do not exceed the performance at a 25% top ratio. Additionally, the audio-visual source localization performance at a 50% top ratio is reduced compared to a 25% top ratio. We analyze that this is due to the introduction of more background or noisy tokens, which are not truly attentive, affecting the region alignment of CrossMAE. Therefore, we ultimately set both the visual and auditory top ratios to 25%.

C. Impact of Mask Strategy and Mask Ratio

CrossMAE uses a masking ratio of 75% for both the visual and auditory modalities, with both images and audio employing a random masking approach. In this section, we investigate the impact of different masking methods and mask ratios on CrossMAE's representational ability.

Firstly, for audio, the model input is a spectrogram of the audio clip, where the two dimensions represent time and frequency domains. Thus, there are four masking strategies for audio: random (unstructured), frequency mask, time mask, and time-frequency mask (structured), as shown in Fig. 2. We evaluated the performance of model variants with different masking methods on retrieval and audio-visual source localization tasks. We find that random masking (unstructured) of the spectrogram performs the best, followed by time-frequency (structured) masking, while masking only time or frequency individually results in a slight performance decrease. We believe that random and time-frequency masking strategies assist the model in reconstruction based on the information surrounding the tokens.

Moreover, we explored the impact of different mask ratios on the visual and auditory modalities. As seen in Fig. 1, a 75% mask ratio is suitable for both modalities. Therefore, we set the visual and auditory masking ratio to 75%.

Table 3. Audio-visual retrieval results on VGGSound.

	Audio → Visual			Visual → Audio		
	R@1	R@5	R@10	R@1	R@5	R@10
(a) CAV-MAE [5]	12.1	31.6	42.4	14.7	35.2	45.9
(b) CL-AV	8.4	2.6	36.0	6.2	18.8	28.8
(c) SelfMAE	2.6	5.6	12.8	2.0	5.4	11.2
(d) CL-SelfMAE	9.6	27.2	39.2	8.2	21.6	38.0
(e) CL-CrossCdtMAE	15.2	42.4	52.8	14.0	41.2	50.8
(f) CL-CrossEmbMAE	19.4	48.6	61.2	18.8	47.2	55.6
(g) CrossMAE (ours)	**22.8**	**56.6**	**70.8**	**21.3**	**54.4**	**60.5**

Table 4. Retrieval results on MSR-VTT dataset.

Variants	audio-visual			visual-audio		
	R@1	R@5	R@10	R@1	R@5	R@10
CAV-MAE [5]	7.6	19.8	30.2	13.3	29.0	40.5
CL-AV	4.8	14.4	28.4	6.2	18.6	33.2
SelfMAE	2.0	9.2	14.8	2.4	9.6	18.8
CL-SelfMAE	5.2	18.6	29.4	7.6	20.4	36.0
CL-CrossCdtMAE	6.8	27.2	40.6	8.0	29.6	41.0
CL-CrossEmbMAE	9.6	28.8	42.4	10.0	33.2	44.8
CrossMAE(ours)	**12.0**	**31.6**	**46.0**	**14.8**	**36.8**	**47.2**

D. More Audio-Visual Retrieval Results

To evaluate the generalization ability of CrossMAE across different datasets more comprehensively, we tested the performance of the model on zero-shot audio-visual retrieval, where the model is pre-trained on AudioSet. Apart from the Flickr-SoundNet dataset in the main text, we provide the performance of audio-visual retrieval on the VGG-SoundSource and MSR-VTT datasets.

MSR-VTT [20] serves as a new large-scale video benchmark for video understanding, particularly the emerging task of translating video to text. It achieves this by gathering 257 popular queries from a commercial video search engine, with 118 videos for each query. In its current ver-

sion, MSR-VTT provides 10,000 web video clips totaling 41.2 hours and 200,000 clip-sentence pairs, covering comprehensive categories and diverse visual content. It represents the largest dataset in terms of sentence and vocabulary size, with each clip annotated with approximately 20 natural sentences.

We compared CrossMAE with the current state-of-the-art (SOTA) audio-visual pre-training method, CAV-MAE, as well as different CrossMAE variants. The results in Table 3 and Table 4 indicate that CrossMAE achieved the best performance, surpassing existing methods by a large margin. We evaluated the audio-visual retrieval performance of CrossMAE. Among the different variants of CrossMAE, we found that CrossMAE had the best retrieval performance, followed by CL-CrossEmbMAE and CL-CrossCdtMAE, CL-SelfMAE, CL-AV and SelfMAE, from which we can draw two conclusions: (i) Cross-Conditioned reconstruction and Cross-Embedding reconstruction are beneficial for retrieval performance, suggesting that region alignment is conducive to global-alignment between the two modalities, promoting retrieval performance. (ii) The training objectives of MAE and contrastive learning are complementary, their combination can promote each other and jointly improve performance.

E. Comparison with Task-Specific Methods of Audio-Visual Source Localization

In the main paper, we evaluate CrossMAE on zero-shot audio-visual source localization without fine-tuning. For fair comparison, we fine-tuned the pre-trained CrossMAE using specific objectives on two datasets. Table 5 shows that without fine-tuning, CrossMAE already outperforms the best result in zero-shot localization performance on the VGG-SoundSource dataset. After fine-tuning, CrossMAE achieves state-of-the-art performance on both datasets, demonstrating its regional ability.

Table 5. Audio-visual source localization results with task-specific models.

Variants	Flickr		VGG-SoundSource	
	CIoU	AUC	CIoU	AUC
EZVSL [13]	72.69	58.70	34.38	37.70
SLAVC [14]	73.84	58.98	39.20	39.46
SSL-TIE [11]	81.50	**61.10**	38.60	39.60
CrossMAE(ours)	**82.40**	59.24	**42.00**	**39.88**

F. Comparison with MAViL

(a)Training strategy. MAViL [9] introduces knowledge distillation and uses teacher-student structure with 2×#parameters to train iteratively, which is orthogonal to our novelty in modeling. **(2)Data augmentation.** MAViL uses strong augmentation like SpecAugment and MixUp during fine-tuning, while we use no augmentation. For fairness, we evaluate CrossMAE with strong augmentation and iterative 2-stage teacher-student. Table 6 shows that just introducing the 2-stage training (+Teacher-student) can achieve comparable performance. Adding strong data augmentation further enhances classification, performing better performance than MAViL. Notably, consistent state-of-the-art performance of joint A-V classification performance underscores CrossMAE's efficacy.

Table 6. Comparison with MAViL.

Variants	A	V	A-V
MAViL [9]	48.7	30.3	53.3
CrossMAE(ours)	47.1	27.2	55.3
+Teacher-student stage	49.0	29.7	57.9
+Strong Augmentation	49.8	30.6	59.1

G. Further Audio-Visual Tasks

We fine-tune the pre-trained model on AVE [18, 19], AVQA [21, 22], and AVS [12, 23] tasks. We adopt their datasets and report evaluation metrics, respectively. Tables show our model can effectively extend to many classic audio-visual tasks and achieve superior or comparable results to SOTA methods, validating CrossMAE's modality interaction and region alignment.

Table 7. AVE.

Method	acc
CMRAN	78.3
CMBS	79.7
Ours	**81.2**

Table 8. AVQA.

Method	acc
Pano-AVQA	66.64
AVQA	69.51
Ours	**71.36**

Table 9. AVS.

Method	mIoU
AVS	78.70
DiffusionAVS	81.38
Ours	**81.86**

H. Limitations

We think there are some limitations of CrossMAE. Firstly, the pre-trained dataset, which consists of 10-second recordings in AudioSet, is relatively short. As a result, it may not adequately learn distant temporal dependencies in audio. Future considerations may include modeling longer audio sequences. Secondly, the dataset scale is limited. AudioSet, which is used by CrossMAE, is around two orders of magnitude smaller than the text corpus used in language counterparts. Moreover, some video samples in the dataset have time periods where sounding objects are not present in the frame; and there are also low-quality videos and mismatched audio-visual pairs, which will introduce noise into the training process. Finally, Future works could also delve into the relationship and optimal schedule and balance of cross-modality reconstruction and contrastive learning.

References

[1] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020. 2

[2] 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. 2

[3] Haytham M Fayek and Anurag Kumar. Large scale audio-visual learning of sounds with weakly labeled data. *arXiv preprint arXiv:2006.01595*, 2020. 1

[4] 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. 2

[5] 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, 3

[6] 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. 2

[7] 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. 2

[8] 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

[9] 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. 4

[10] 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. 1

[11] 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. 4

[12] 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. 4

[13] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. *arXiv preprint arXiv:2203.09324*, 2022. 2, 4

[14] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. In *Advances in Neural Information Processing Systems*, 2022. 2, 4

[15] 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. 1

[16] 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. 2

[17] 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. 2

[18] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19989–19998, 2022. 4

[19] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *ACM International Conference on Multimedia*, 2020. 4

[20] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 3

[21] 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. 4

[22] 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. 4

[23] 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. 2, 4