

Article

A Lightweight Framework for Audio-Visual Segmentation with an Audio-Guided Space–Time Memory Network

Yunpeng Zuo and Yunwei Zhang *

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China; yunpengzuo@126.com

* Correspondence: zhangyunwei72@gmail.com

Abstract: As a multimodal fusion task, audio-visual segmentation (AVS) aims to locate sounding objects at the pixel level within a given image. This capability holds significant importance and practical value in applications such as intelligent surveillance, multimedia content analysis, and human–robot interaction. However, existing AVS models typically feature complex architectures, require a large number of parameters, and are challenging to deploy on embedded platforms. Furthermore, these models often lack integration with object tracking mechanisms and fail to address the issue of the mis-segmentation of unvoiced objects caused by environmental noise in real-world scenarios. To address these challenges, this research proposes a lightweight audio-visual segmentation framework incorporating an audio-guided space–time memory network (AG-STMNet). First, a mask generator with a scoring mechanism was developed to identify sounding objects from generated masks. This component integrates FastSAM, a lightweight, pre-trained, object-aware segmentation model, with WAV2CLIP, a parameter-efficient audio-visual alignment model. Subsequently, AG-STMNet, an audio-guided video object segmentation network, was introduced to track sounding objects using video object segmentation techniques while mitigating environmental noise. Finally, the mask generator and AG-STMNet were combined to form the complete framework. The experimental results demonstrate that the framework achieves a mean Intersection over Union (mIoU) score of 41.5, indicating its potential as a viable lightweight solution for practical applications.



Academic Editors: Douglas O'Shaughnessy, Yi Wang and Seong Hyeon Hong

Received: 6 April 2025

Revised: 19 May 2025

Accepted: 10 June 2025

Published: 11 June 2025

Citation: Zuo, Y.; Zhang, Y. A

Lightweight Framework for Audio-Visual Segmentation with an Audio-Guided Space–Time Memory Network. *Appl. Sci.* **2025**, *15*, 6585. <https://doi.org/10.3390/app15126585>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: lightweight; multimodal fusion; audio-visual segmentation (AVS); video object segmentation (VOS); space–time memory (STM) network

1. Introduction

Audio-visual segmentation (AVS), as a multimodal fusion task, intends to identify the positions of sounding objects at the pixel level in a given image [1]. It builds on Sound Source Localization (SSL), which integrates audio and image information to locate sounding objects in an image and mark their regions with a heatmap [2]. Extending SSL, Zhou et al. introduced the AVS task and the corresponding AVSbench dataset. This task refines SSL by generating precise masks of sound-emitting objects, enabling accurate localization and boundary delineation. AVS holds significant value in both research and application. On the one hand, AVS integrates the auditory modality with the conventional visual modality to improve the fault tolerance and perception capabilities of the system. On the other hand, AVS technology has broad application potential in domains such as home assistance, smart home control, security monitoring, and healthcare assistance. For example, in security monitoring, fusing visual and acoustic information enables more effective identification

and tracking of abnormal noise sources within a given area. This extends the surveillance range, improves monitoring efficiency, and serves as a valuable supplement to fixed video surveillance systems. Due to its practical significance, AVS has attracted substantial research interest and emerged as a prominent topic in the field.

Researchers have significantly enhanced the effectiveness of AVS by leveraging the potential of audio-visual data. For instance, Liu et al. proposed a framework that utilizes unlabeled frames, improving marginal performance gains by incorporating motion cues from adjacent frames and semantic information from distant frames [3]. Similarly, Chen et al. introduced an approach that strengthens audio cues through a Bidirectional Audio-Video Decoder (BAVD), enabling continuous interaction between audio and video modalities [4]. Guo et al. developed a hierarchical encoder collaboration module and employed neural architecture search to optimize information interaction [5]. Several studies have focused on the bilateral interaction between visual and auditory modalities. Hao et al. proposed a bidirectional generation framework where visual-to-audio projection reconstructs audio features from object segmentation masks while minimizing reconstruction errors [6]. Yang et al. designed a transformer-based multi-order bilateral architecture, Cooperation of Multi-Order Bilateral Relations (COMBO), which includes a Bilateral Fusion Module (BFM) to precisely align visual and auditory signals [7]. Mao et al. addressed AVS through an Explicit Conditional Multimodal Variational Auto-Encoder (ECMVAE), emphasizing effective representation learning [8]. Other researchers have emphasized enhancing semantic associations in AVS. Liu et al. developed a method based on cross-modal semantic interaction to address ambiguity caused by reliance on visual saliency detection [9]. Wang et al. introduced a text semantic-guided technique that leverages textual cues within visual scenes to strengthen audio semantics [10]. Additionally, Wang et al. proposed the Progressive Confident Masking Attention Network (PMCANet), which employs an attention mechanism to enhance semantic awareness through query token selection [11].

With the advent of the Segment Anything Model (SAM), a universal pre-trained image segmentation model, numerous methods have leveraged this approach for AVS due to its high accuracy and flexible interface [12]. For instance, Mo et al. introduced an efficient audio-visual localization and segmentation framework that integrates visual features with audio features extracted from the SAM's image encoder to construct pixel-level cross-modal representations [13]. Huang et al. proposed a training-free framework, Audio-Language-Referenced SAM 2 (AL-Ref-SAM2), built on SAM2 and a GPT. This framework utilizes GroundingDINO for single-frame object recognition, GPT-4 for selecting reference information, and SAM2 for segmenting recognized objects across video sequences [14]. Similarly, Bhosale et al. introduced Modality Correspondence Alignment (MoCA), an AVS approach that seamlessly integrates pre-trained foundational models such as DINO, the SAM, and ImageBind. Leveraging the complementary knowledge of these models, MoCA optimizes their synergistic application to achieve multimodal correlation [15].

In pursuit of higher accuracy, conventional AVS models often employ complex architectures, resulting in insufficient efficiency. To address this issue, Liu et al. introduced SAMA-AVS, a lightweight model that utilizes the pre-trained SAM for AVS tasks. By integrating a limited number of trainable parameters and adapters, this approach effectively achieves audio-visual fusion and interaction during the encoding phase [16]. Similarly, Nguyen et al. developed Segment Audio-Visual Easy Way (SAVE), a lightweight AVS model that fine-tunes transformer-based image encoders and employs a residual audio encoder to generate sparse audio feature hints, facilitating effective audio-visual fusion [17]. Lin et al. proposed a Vision Transformer (ViT)-based lightweight AVS framework that injects a minimal number of trainable parameters into each ViT layer for fine-tuning, adapting pre-trained ViTs for audio-visual tasks [18]. Additionally, Xu et al. introduced Each

Performs Its Functions (PIF), an efficient AVS method that divides the task into two stages: relevance learning and segmentation refinement. This approach leverages deep features for cross-modal interaction and shallow features to enhance segmentation results [19].

However, despite these advancements, the aforementioned algorithms still incorporate a significant number of parameters, and the SAM exhibits high computational complexity and extensive processing time. These limitations pose challenges for deployment on embedded platforms, underscoring the need for further improvements in model efficiency.

In addition to the challenge of model lightweighting, general AVS methods often fail to integrate object tracking and neglect the mis-segmentation of unvoiced objects caused by environmental noise. These limitations remain critical barriers to practical application, as illustrated in Figure 1.

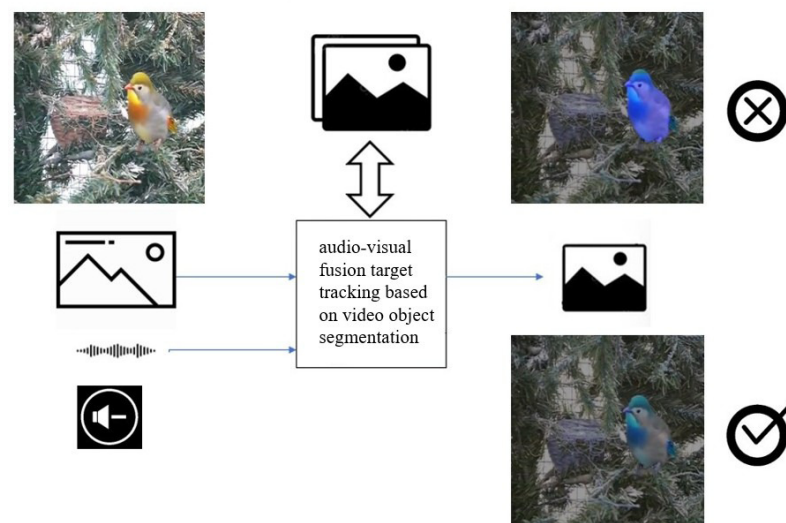


Figure 1. In real-world scenarios, the phenomenon of the mis-segmentation of unvoiced objects often occurs as a result of environmental noise. When conducting object segmentation tasks, especially for those targets that are unvoiced, the presence of environmental noise can severely disrupt the normal operation of the segmentation model. Moreover, directly applying the audio-visual segmentation algorithm to segment each frame of an image may result in an unstable target mask, limiting its practical applicability in real-world scenarios. To address this, integrating a tracking mechanism allows the model to utilize historical segmentation results, enabling it to track vocalized objects more effectively. This approach not only improves the accuracy of the outputs but also enhances their stability.

To address the above issues, this study proposes a lightweight AVS framework that incorporates an audio-guided space–time memory network. The overall framework diagram of the proposed architecture is shown as Figure 2. The key contributions of this study are divided into the following two aspects.

- To achieve lightweight audio-visual segmentation, this research proposes an efficient audio-visual mask generator leveraging pre-trained lightweight models, Fastsam and Wav2CLIP. Additionally, a mask-scoring mechanism is designed to optimize the mask, thereby enhancing the accuracy and precision of target segmentation;
- In addition to the audio-visual fusion mask, to enable continuous tracking of the vocal target guided by the fusion mask, this research proposes a video object tracking network based on a space–time memory network (STM-Net). The audio feature extraction and fusion module within the network detects changes in the audio signal and transmits the output information to the upsampling module of the mask decoder to guide the generation of the final segmentation mask, thereby suppressing the interference of environmental noise.

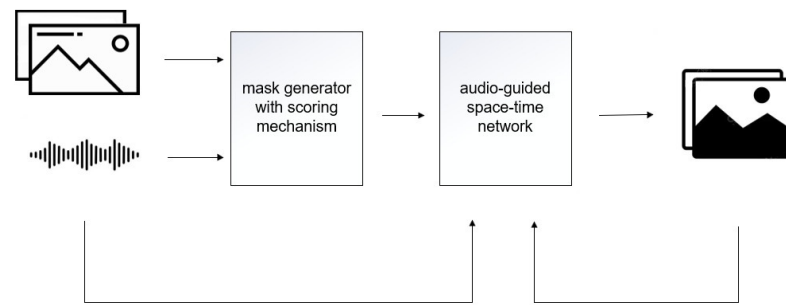


Figure 2. The proposed framework consists of two components: the audio-visual mask generator with a scoring mechanism and the audio-guided space–time memory network. First, paired images and audio clips are input into the mask generator. Subsequently, this module generates the mask of a sounding object and transmits it to the audio-guided space–time memory network. Lastly, the audio-guided space–time memory network leverages both historical and current masks to track the sounding object and determine whether to output the target mask based on the audio guidance.

Both of the two aspects will be elaborated on separately in the Methodology Section. The combination of these two aspects constitutes the overall framework proposed in this research. In the Experiments and Results Section. Experiments for the two modules were conducted separately, and the experimental verification of the overall framework was incorporated into the experiments of the second module.

2. Methodology

2.1. Mask Generator

The mask generator consists of three interconnected modules: a lightweight image segmentation model (Fastsam), an audio-visual modal alignment model (Wav2CLIP), and a mask-scoring module. The Fastsam model is a lightweight segmentation model that achieves performance comparable to the SAM while attaining a running speed more than 50 times faster. It addresses the issue of high computational costs associated with the practical application of the SAM, enabling deployment on embedded platforms [20]. The Wav2CLIP model achieves robust audio-visual modality alignment. It is lightweight, containing fewer than 50M parameters, which makes it well-suited for deployment on embedded platforms with constrained computational resources. Additionally, it can generate frame-level audio embeddings along the temporal axis, offering significant advantages for tasks requiring precise frame-level embeddings and enabling adaptive inference in video-related applications [21]. The lightweight architecture of the Fastsam and Wav2CLIP models, combined with the simplicity and efficiency of the mask-scoring mechanism, ensures that the mask generator achieves high parameter efficiency while making the system scalable.

However, the preliminary mask produced by the image segmentation model demonstrates fine-grained segmentation, dividing the target object into multiple parts. This outcome stems from the pre-trained model’s ability to learn diverse object characteristics and knowledge from extensive datasets, enhancing its versatility in recognizing and generating masks for various objects. Excessively detailed segmentation not only complicates mask selection but also undermines the efficacy of correlation calculations in the audio-visual alignment model.

To mitigate these challenges, two strategies can be implemented. The first approach involves enlarging the feature map in the model’s detection head. This method is particularly effective in scenarios where the target mask size remains relatively stable, as it simplifies the problem and accelerates processing speed, albeit at the risk of overlooking smaller targets. The second approach focuses on filtering and stitching the output masks. While this method incurs higher computational costs and segments all target types, it offers more

comprehensive information and is better suited for scenarios with substantial variations in target size.

In most application scenarios, small targets are less frequent but tend to be more conspicuous, easier to observe, and meaningful when they do appear. Conversely, large targets, which are more common, are typically not divided into numerous parts. Therefore, this study adopts a second approach and introduces a mask-scoring mechanism to streamline the mask screening and stitching process. Specifically, the mask-scoring mechanism utilizes the full-instance mask segmentation map obtained from the Fastsam model to select and combine masks, thereby optimizing the generated masks and improving the accuracy and precision of object segmentation.

As illustrated in Figure 3, the proposed mask-scoring module involves two critical processes. The first process calculates the correlation between audio and visual embeddings, which forms the foundation of audio-visual fusion. Specifically, this step identifies the mask that most accurately corresponds to the current input audio. As the Wav2CLIP model aligns visual and audio embeddings within the same space, the relevance of audio and visual information is determined by computing the normalized cosine similarity between the embeddings. The second process derives the final mask based on the scoring formula, which can be mathematically represented by the following formula:

$$U = \bigcup_{i=1}^n x_{\delta(i)} \quad (1)$$

$$\delta : \forall j, k \in \{1, \dots, m\}, j < k \Rightarrow s(x_{\delta(j)}) \geq s(x_{\delta(k)}) \quad (2)$$

$$s(x_i) = \lambda_1 * S_i + \lambda_2 * A_i \cdot V_i \quad (3)$$

In Equation (1), n indicates the total number of masks, x_i denotes each image block, and δ represents the ranking function; the index of the ranking corresponds to the result computed by the scoring function, as presented in Equation (2). In Equation (3), λ_1 and λ_2 denote the weighting parameters for area and similarity, respectively, S_i represents the normalized area, and A_i and V_i correspond to the audio embedding and image embedding.

The specific processing flow of the mask-scoring module is structured as follows: First, the cosine similarity between the video embedding and the audio embedding is computed, with the results normalized and ranked in descending order. Second, the area of each image block is calculated, normalized, and sorted in descending order. These normalized similarity and area values are then assigned different weights, summed, and ranked in descending order to produce the final scores. Finally, the n masks with the highest scores are selected and concatenated to generate the final mask.

In order to determine the optimal combination of the parameters λ_1 , λ_2 , and n for the mask-scoring mechanism, we iteratively fine-tune these parameter values based on the output mask map until optimal results are achieved. Specifically, the parameter n is assigned values of 3, 4, and 5 across three testing rounds. In each round, the variable λ_1 begins at 0 and incrementally increases to 1 in steps of 0.1. Given that $\lambda_1 + \lambda_2 = 1$, the variable λ_2 consequently starts at 1 and decrementally decreases to 0 in steps of 0.1.

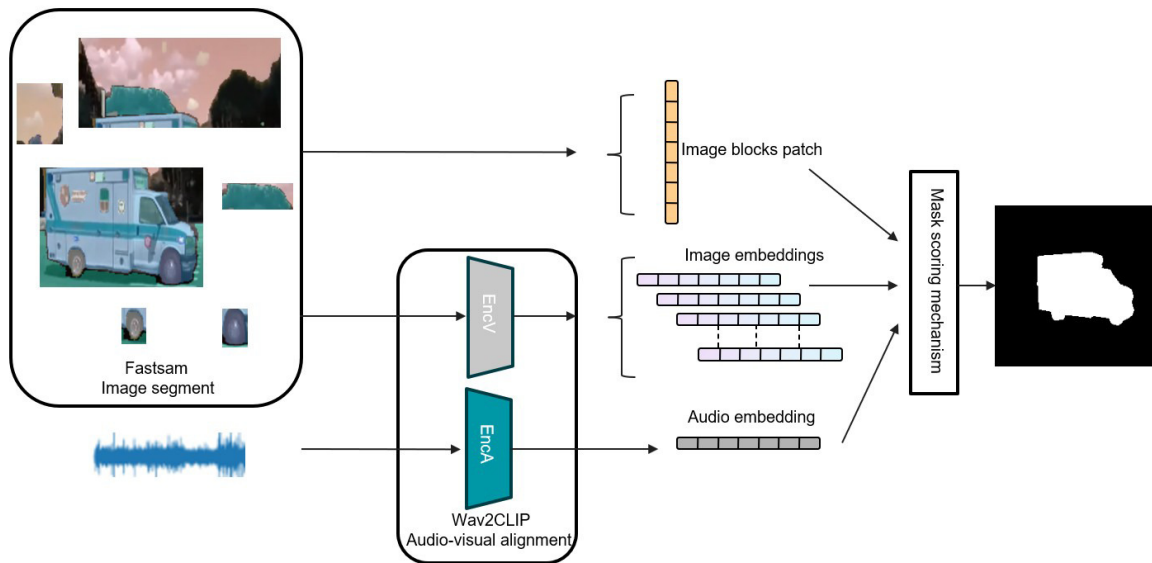


Figure 3. The mask generator operates through the following detailed procedure. First, an image is input into FastSam to generate a preliminary segmentation mask. Second, the image is cropped based on the mask to extract the candidate region's image block. Next, the image block and its synchronized audio segment are processed by the Wav2CLIP model to produce corresponding image embeddings and an audio embedding. Finally, the image embeddings, audio embedding, and image block patches are fed into the mask-scoring module, where the final mask is generated according to the predefined scoring mechanism.

2.2. Audio-Guided Space–Time Memory Network

Space–time memory networks excel at tracking targets using given segmentation masks [22]. Unlike semi-supervised video object segmentation methods, their key advantage is the ability to store and retrieve historical information, which both speeds up the process and improves tracking accuracy. If the space–time memory network model is sufficiently lightweight, it can achieve real-time target tracking.

However, beyond the challenges of tracking accuracy and speed, practical application environments introduce additional complexities. Ubiquitous environmental noise and interfering sound sources can lead the model to segment silent targets. To address this issue, this study proposes an audio-guided space–time memory network (AG-STMNet), as illustrated in Figure 4. The network consists of two key components: the space–time memory network, which conducts video object segmentation guided by the fusion mask, and the audio feature extraction and fusion module.

In the space–time memory network module, the historical frame object masks are encoded by the memory encoder, generating corresponding keys and values. These keys and values are then chronologically concatenated and converted to the memory key embedding k^M and memory value embedding v^M . Upon the arrival of the current frame, the query encoder is employed to encode it, producing the corresponding key embedding k^Q and value embedding v^Q . Subsequently, a spatial matching mechanism is utilized to determine whether each pixel belongs to the foreground object, thereby completing the segmentation of the target object in the current frame. This process effectively computes the spatio-temporal attention of each pixel in the query image with respect to the pixels in the historical frames. The spatio-temporal memory query process is mathematically expressed by the following formula [22]:

$$y_i = [v_i^Q, \frac{1}{Z} \sum_j f(k_i^Q, k_j^M) v_j^M] \quad (4)$$

y represents the query output, where i and j denote the indices of each position in the query embedding and memory embedding feature map, respectively. The notation $[\cdot, \cdot]$ indicates the concatenation and stacking operation, while f denotes the similarity calculation function.

$$f(k_i^Q, k_j^M) = \exp(k_i^Q \circ k_j^M) \quad (5)$$

The $\circ$ operation represents the dot product operation, and Z denotes the normalization factor.

$$Z = \sum_j f(k_i^Q, k_j^M) \quad (6)$$

In the aforementioned formula, $k^Q \in \mathbb{R}^{H \times W \times C/8}$, $v^Q \in \mathbb{R}^{H \times W \times C/2}$, $k^M \in \mathbb{R}^{T \times H \times W \times C/8}$, $v^M \in \mathbb{R}^{T \times H \times W \times C/2}$, where H denotes the height of the image, W represents the width of the image, C signifies the number of dimensions in the feature map, and T corresponds to the number of historical frames retained.

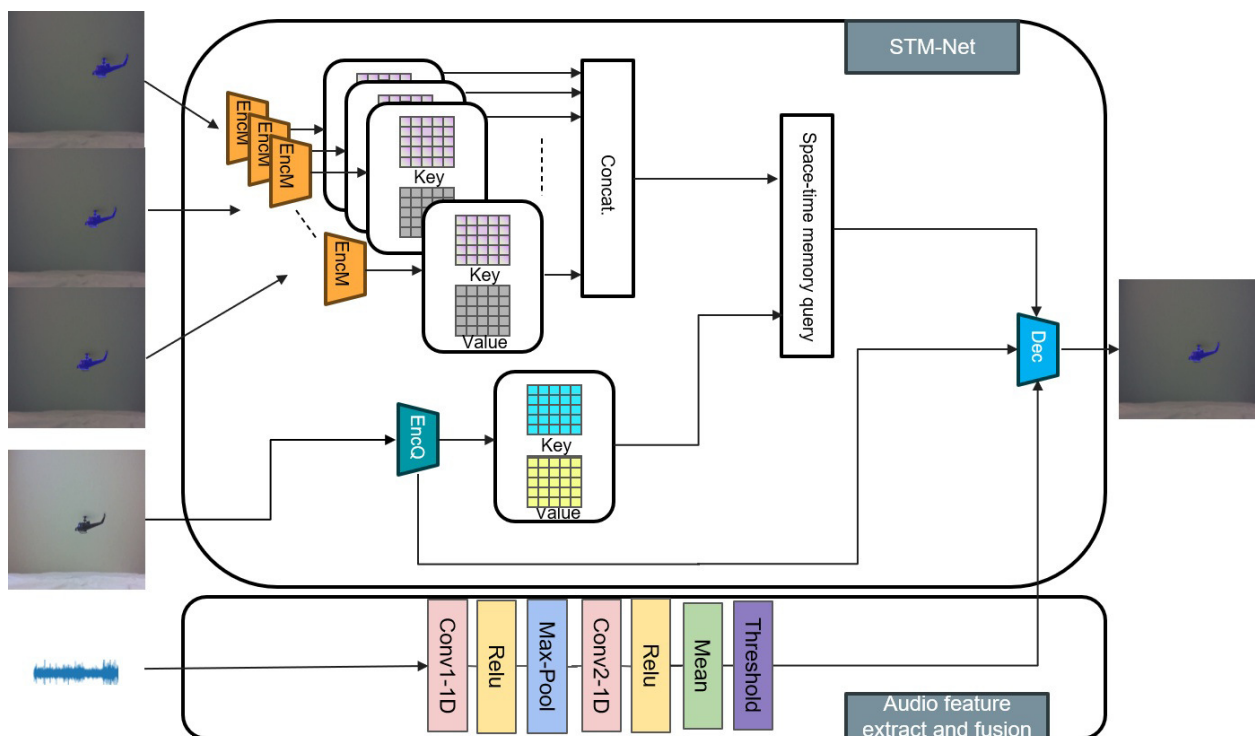


Figure 4. The audio-guided space–time memory network (AG-STNet) incorporates an audio feature extraction fusion module and a video object-segmentation module based on a space–time memory network. The audio feature extraction fusion module processes synchronized audio frame to extract sound features and detect whether the sound has disappeared or changed. The video object-segmentation module based on a space–time memory network encodes the current image query by leveraging the previous fusion mask and stored historical masks. It then integrates the results from the audio feature extraction and fusion module to the decoder to output the current mask, thereby enabling suppression of environmental noise.

The audio feature extraction and fusion module consists of two one-dimensional convolutional layers. The first layer maps the input dimension to an intermediate dimension, followed by dimensionality reduction via max pooling. The second one-dimensional convolutional layer then maps the intermediate dimension to the output dimension, extracting the final audio feature. The specific dimensional configurations for each layer are provided in Table 1.

The 1D convolutional layer processes the input sequence by applying a convolution kernel, which captures localized patterns within the sequence. Specifically, the 1D con-

volution kernel detects variations in frequency and pitch intensity in the audio signal. The resulting features undergo dimensionality reduction and aggregation, with the global mean value representing the overall audio intensity. This value is compared to a predefined threshold to determine whether the mask decoder output is activated. If the global mean exceeds the threshold, the decoder output is enabled; otherwise, a full background output is generated. This switching mechanism, embedded within the audio feature extraction and fusion module, effectively reduces the impact of environmental noise and interfering sound sources.

Table 1. Parameters of the audio feature extraction and fusion module.

Layer Type	Kernel Size	Step Size	Kernel Size	Input Shape	Output Shape
Conv1-1d	3	1	1	$1 \times 22,050$	$64 \times 22,050$
MaxPool1d	2	2	-	$64 \times 22,050$	$64 \times 11,025$
Conv2-1d	3	1	1	$64 \times 11,025$	$128 \times 11,025$

3. Experiments and Results

3.1. Mask Generator

Based on the structure and process outlined in the previous section, the mask generator was constructed. Experiments were subsequently conducted to evaluate and validate the mask-generator method, as well as to assess the effectiveness of the mask-score selection mechanism.

3.1.1. Dataset, Evaluation Metrics, and Hardware

The AVSbench dataset, specifically designed for fine-grained audio-visual segmentation tasks, was employed in the experiments for the mask generator due to its compact size and ease of use. Each video in the dataset was trimmed to a 5 s duration, with the last frame from every 1 s interval extracted as an image frame. The test subsets included masks for all image frames to streamline testing and validation [1].

Since the proposed method focuses exclusively on single-object segmentation, only the single-source subset (S4) of AVSbench was utilized. Furthermore, as the Wav2CLIP model relies on frame-level audio embedding, the dataset's audio log MEL spectrogram was not used. Instead, segments extracted directly from the original audio files served as the audio input.

The evaluation metrics included the Jaccard Index, mean Intersection over Union (mIoU), and F-Score, which are widely accepted as critical measures in audio-visual segmentation. The experiments were conducted on a laptop equipped with an Intel i9-12900H CPU, 32 GB RAM, and an NVIDIA RTX 3080 GPU with 16 GB VRAM. The software environment consisted of Ubuntu 20.04, Python 3.8.16, and PyTorch 2.4.1.

3.1.2. Experiments

(1) Parameter determination

We define the parameters λ_1 , λ_2 , and n for the mask-scoring mechanism used in model inference. As stated in the previous section, λ_1 and λ_2 denote the weighting parameters for area and similarity, respectively, and n indicates the total number of masks.

Table 2 presents the typical parameter combinations and their corresponding indices derived from the experiments. The optimal combination of parameters is $\lambda_1 = 0.1$, $\lambda_2 = 0.9$, $n = 3$; it is evident that the mask-scoring mechanism can enhance the performance of object segmentation. Due to the utilization of the FastSAM model based on CNN object detection networks, the overall recognition performance is superior. The masks of objects

are more complete, the number of generated masks is reduced, and the area weight is smaller than the similarity weight.

Table 2. Parameters of the mask-scoring mechanism.

λ_1	λ_2	n	mIoU	F-Score
0.1	0.9	3	54.5	0.565
0.1	0.9	4	42.4	0.435
0.1	0.9	5	41.3	0.364
0.5	0.5	3	32.4	0.367
0.9	0.1	3	19.2	0.238

(2) Comparison experiments

The ablation study was performed first. The parameter λ_1 in the mask-scoring mechanism was set to 0, while λ_2 was set to 1. This implies that only the audio-visual embedding correlation based on the Wav2CLIP model is considered. Additionally, the parameter n was set to 1, meaning that only the mask with the highest correlation to the audio was output, and no mask concatenation was performed. Comparing the experimental metrics obtained without using the mask-scoring mechanism to those produced by the model within its optimal parameter configuration, we analyzed the impact of the mask-scoring mechanism on the model's performance.

Subsequent to the ablation study, the performance metrics of the proposed mask generator were compared against those of other audio-visual segmentation methods, yielding the results presented in Table 3. Herein, TPAVI-R50 is the reference benchmark method for the AVSbench dataset [1]. SAM zero-shot represents the zero-shot learning approach leveraging the SAM and Ground-DINO models described in [23]. AV-SAM denotes the fine-tuning technique utilizing the SAM model as outlined in [13]. AL-Ref-SAM 2 refers to the method introduced in [14], which employs the SAM model for mask generation and the GPT-4 model for mask selection.

Table 3. Comparison experiments.

	mIoU	F-Score
<i>Fastsam</i>	41.7	0.526
<i>Fastsam + scoring mechanism</i>	54.5	0.565
<i>TPAVI – R50</i>	72.6	0.482
<i>SAM zero – shot</i>	51.8	0.626
<i>AV – SAM[14]</i>	40.5	0.568
<i>AL – Ref – SAM2[16]</i>	70.5	0.811

The data in the table can be interpreted as follows: (1) The module incorporating the mask-scoring selection mechanism shows a significant improvement in the output index, validating its effectiveness. (2) The proposed mask generator outperforms the AV-SAM method, which relies on SAM model fine-tuning, indicating that it can serve as a lightweight alternative to SAM fine-tuning. (3) The proposed method exceeds the SAM zero-shot method in terms of the mean mIoU metric, demonstrating the efficacy of the introduced mask-scoring mechanism in enhancing output mask accuracy. (4) While the TPAVI-R50 benchmark and the AL-Ref-SAM 2 method (which integrates SAM and GPT-4 models) achieve higher accuracy and precision in object segmentation on the AVSbench dataset, the proposed method is more computationally efficient. It is better suited for platforms with limited resources due to its reduced dependence on heavy feature extraction networks like ResNet50 and advanced language models like GPT-4.

(3) Visualization and case study

To verify the effectiveness of the proposed method, the typical segmentation results obtained during testing were plotted and compared with the labeled mask maps provided by the dataset.

In Figure 5, it is evident that the mask generator effectively produces the target mask in the image based on audio prompts. Although the output accuracy is slightly lower compared to that with the real mask, the performance remains within an acceptable range. As a result, the proposed mask generator provides a concise and efficient solution for the practical implementation of audio-visual segmentation tasks, eliminating the need for additional model training and significantly reducing computational costs.

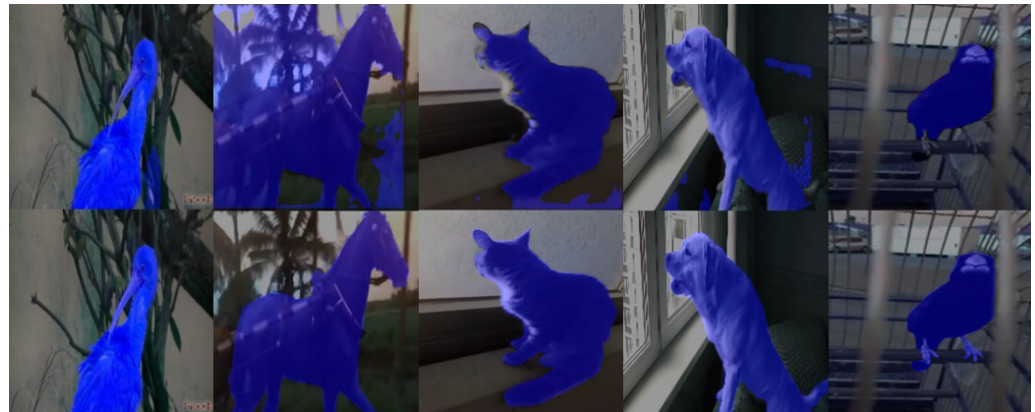


Figure 5. The visualization is presented above. The top row presents the mask overlay images output by the model, while the bottom row displays the images annotated with overlay masks.

3.2. Audio-Guided Space–Time Memory Network

In this section, based on the network architecture and process outlined in the preceding section, the construction of the audio-guided space–time memory network (AG-STMNet) is described. Following the training of the network, it was rigorously tested and validated. Subsequently, the mask generator and AG-STMNet were integrated into a complete target tracking framework to evaluate the overall performance of the proposed method.

3.2.1. Dataset, Evaluation Metrics, and Hardware

AVSbench is the first publicly released and most widely utilized audio-visual segmentation dataset. Ref-AVSbench represents a more recent and extensive dataset for audio-visual segmentation and possesses two significant characteristics: (1) It contains more audio and video frames and clips. The authors segmented each 10 s video into 10 equal 1 s clips, extracted the first frame of each clip as an image frame, and generated corresponding binary mask annotations. (2) An empty subset is introduced, where the target is silent and the associated mask labels are also empty. The extensive number of data frames in the Ref-AVS Bench dataset renders it particularly suitable for learning temporal information. Furthermore, the introduction of the empty subset enables models trained on this dataset to mitigate overfitting, suppress environmental noise, and reduce the mis-segmentation of unvoiced objects. Because of these characteristics, the Ref-AVSbench dataset was utilized for training in the experiments described in this section.

AVSbench is the first publicly released and most widely used audio-visual segmentation dataset. Ref-AVSbench, a more recent and comprehensive dataset, offers two key features: (1) It includes a larger number of audio and video frames and clips. Each 10 s video is segmented into 10 equal 1 s clips, with the first frame of each clip extracted as an image frame and accompanied by binary mask annotations. (2) It introduces an empty

subset where the target is silent, and the associated mask labels are also empty. The extensive data frames in Ref-AVSbench make it particularly suitable for learning temporal information, while the empty subset helps mitigate overfitting, suppress environmental noise, and reduce the mis-segmentation of unvoiced objects. Therefore, Ref-AVSbench was utilized for training in this study.

Since the proposed method focuses solely on single-target tracking, when utilizing the Ref-AVSbench dataset, the data are preprocessed via sorting and filtering to retain only the single-source case. Additionally, it should be noted that while the Ref-AVSbench dataset includes an empty subset, most instances within the single-source subset are continuously voiced. Correspondingly, mask annotations predominantly label the target foreground for each segment. To enhance the audio-feature-extraction fusion module's sensitivity to changes in frequency and pitch intensity in the audio, as well as to mitigate the impact of environmental noise and interfering sound sources on target segmentation, modifications to the dataset are necessary. Specifically, certain audio segments are extracted and muted, with their corresponding mask labels designated as the background. Therefore, through these adjustments, learning samples capturing transitions from sounding to silence are incorporated, and the model's robustness with regard to its ability to recognize environmental sound variations is improved.

The metrics employed in the experiments include the Jaccard Index, mean Intersection over Union (mIoU), and F-Score, which are the same as those used for evaluating the mask generator. A cloud server was used for the experiment. It is equipped with an AMD EPYC 9354 CPU, RAM of 64.4 GB, and an NVIDIA RTX 4090 GPU with VRAM of 24 GB. The operating system is Ubuntu 22.04, and the software environment includes Python 3.11.16, PyTorch 2.2.2, and Docker 26.1.0.

3.2.2. Experiments

(1) Training

To enhance the accuracy and generalization capability of AG-STMNet while reducing its parameter count for greater lightweight efficiency, a two-stage training method was employed in this experiment.

In the first stage, ResNet18 was used as the backbone network, and the space–time memory network component was trained on the MS COCO dataset. Subsequently, the encoder parameters of the trained space–time memory network were frozen, and an audio feature extraction and fusion module was integrated into the decoder. Fine-tuning was then performed using the Ref-AVSbench dataset.

During the initial training of the space–time memory network, video sequences were processed frame-by-frame in groups of three frames [22]. The detailed procedure is as follows: (1) Key–value pairs are extracted from the first frame. (2) The second frame uses the memory prediction mask from the first frame to generate the prediction probability for the current frame. After calculating the loss for the second frame, the memory is updated. (3) The third frame integrates the memories of the preceding two frames for prediction, and the loss for the third frame is computed. The total loss is obtained by summing the prediction losses of the second and third frames, with the loss formula as follows:

$$\text{loss} = n2_loss + n3_loss = L(n2_logit, n2_label) + L(n3_logit, n3_label) \quad (7)$$

$$L(p, y) = -[y * \log(p) + (1 - y) * \log(1 - p)] \quad (8)$$

Let L denote the binary cross-entropy loss function, where p represents the predicted probability and y denotes the label value.

In the second training stage, the audio feature extraction fusion module was constructed and integrated into the AG-STMNet architecture alongside the space–time memory network module. To ensure training stability, the parameters of the STMNet memory encoder and query encoder were frozen. The Ref-AVSbench dataset was used for training, with binary cross-entropy loss as the model’s loss function. The space–time memory network was configured to store three historical frames, and the threshold value for the audio feature extraction fusion module was set to 0.5. Additionally, the label mask of the first frame in the dataset was used as initial input.

(2) Testing and Evaluation

(1) Audio Feature Extraction and Fusion Module Test: The Ref-AVSbench test dataset was imported into the trained network. Consistent with the training setup, the space–time memory network stored three historical frames, the audio feature extraction and fusion module threshold was set to 0.5, and the label mask from the first frame was used as the model input. The network’s performance was evaluated based on the mIoU and F-score of the model output. (2) Audio Feature Extraction and Fusion Module Ablation Experiment: The audio module was removed from AG-STMNet, and only the image and mask data from the Ref-AVSbench test dataset were used for evaluation. The testing procedure remained identical to the previous setup, and the impact of the audio module on the model’s performance was analyzed using the output metrics. (3) Complete Model Test: The mask generator was integrated with AG-STMNet to form a complete model. The Ref-AVSbench test dataset was imported to evaluate the complete model’s performance. During testing, the output of the audio-visual fusion mask generator served as the input to AG-STMNet, while all other settings remained unchanged. The experimental results are presented in the table below:

Based on the data presented in Table 4, the following interpretations can be drawn regarding the model performance:

Table 4. Testing results of AG-STMNet.

STM-Net	Audio-Guided Module	Mask Generator	mIoU	F-Score
✓			71.7	0.725
✓	✓		53.2	0.551
✓	✓	✓	41.5	0.535

(a) STMNet without the Audio Module: The model achieves an mIoU score exceeding 70%, which substantiates the effectiveness of the trained network architecture. This baseline configuration demonstrates robust capability in generating accurate segmentation masks for the current frame by leveraging historical information and image queries. The network’s ability to perform simultaneous image segmentation and target tracking, guided by the fusion mask, establishes a strong foundation for subsequent module integration.

(b) Incorporation of the Audio-Guided Module: The integration of the audio-guided module results in a marginal decrease in the model’s performance metrics. This slight reduction, while statistically insignificant, suggests that the AG-STMNet configuration maintains operational effectiveness in most straightforward environmental scenarios. The model successfully incorporates audio information for target segmentation, albeit with minor limitations. The observed performance degradation can be attributed to instances of mis-masking, which originate from the audio module’s current structural simplicity. These limitations present opportunities for enhancement through more sophisticated architectural refinements and improved audio-visual feature fusion mechanisms.

(c) Complete Model Configuration: The comprehensive model, while exhibiting a relatively modest performance index compared to individual module configurations, demonstrates consistent and reasonable target segmentation efficacy. This validation confirms both the operational effectiveness of the complete model architecture and the fundamental feasibility of the proposed methodology. The comparative reduction in performance metrics primarily stems from error propagation during model inference processes, particularly in the integration of multimodal information streams. This observation underscores the necessity for further optimization in the information fusion pipeline and error correction mechanisms within the complete model architecture.

(3) Visualization and case study

To validate the effectiveness of the proposed method, a representative sample instance was selected from the dataset, where the target transitions from sounding to silent. This example was input into the trained model for inference, and the segmentation results were visualized. These results were then compared with the labeled mask map provided by the dataset, confirming the audio module's effectiveness in suppressing environmental noise.

As shown in Figure 6, when the target sound is absent in the third frame, the model does not segment the target and instead outputs the background. This indicates that the model can dynamically adjust the mask output in response to changes in the audio input, effectively suppressing environmental noise and interfering sound sources. In practical applications, the threshold of the audio module can be flexibly adjusted to enhance noise suppression and ensure adaptability to diverse acoustic environments.

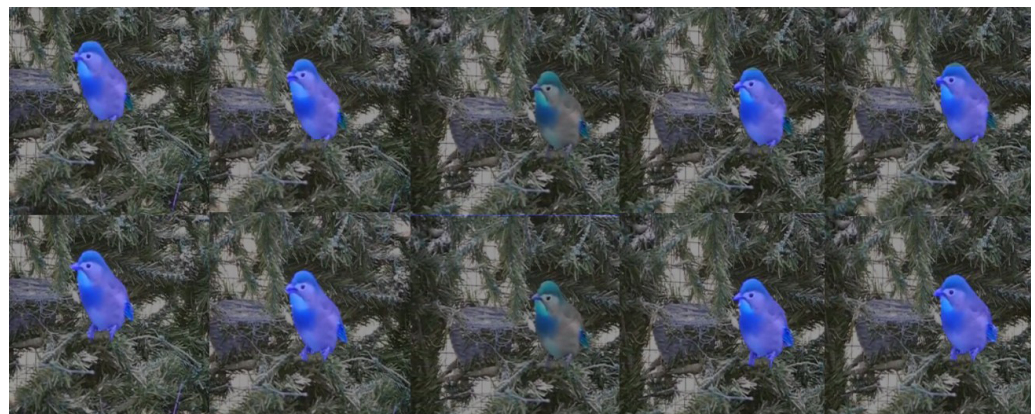


Figure 6. The visualization is presented above. The top row presents the mask overlay images output by the model, while the bottom row displays the images annotated with overlay masks.

4. Conclusions and Discussion

In this research, we proposed a lightweight audio-visual segmentation method. This method comprises two key components: the first is an audio-visual fusion mask generator, and the second is an audio-guided video object-segmentation network based on space-time memory networks. By integrating these two components, the proposed method enables the completion of audio-visual segmentation for application scenarios as follows:

- (1) To achieve lightweight audio-visual fusion object segmentation, a mask generator based on a pre-trained audio-visual modal alignment model is proposed. The pre-trained lightweight image segmentation model, FastSAM, is employed to generate initial masks. A mask-scoring mechanism is then designed to guide the refinement of the final mask output, thereby realizing audio-visual target mask generation. On the AVSbench test dataset, the proposed mask generator achieves an mIoU score of 54.5.

- (2) In order to achieve target tracking with audio-visual fusion mask prompts and address the issue of model mis-segmentation caused by environmental noise, a lightweight sound-controlled target tracking network is proposed. A space–time memory network is employed for target tracking and segmentation, while an audio feature extraction and fusion module is introduced to suppress environmental noise and interference sound sources. The experimental results demonstrate that the proposed method can effectively track targets under audio-visual fusion mask guidance, dynamically adjust the mask output in response to audio changes, and successfully suppress environmental noise. This network achieved an mIoU score of 53.2 on the Ref-AVSbench (S4) test dataset.
- (3) To evaluate the complete model, the mask generator was integrated with the audio-guided space–time memory network. The experimental results on the Ref-AVSbench (S4) test dataset demonstrate that the complete model is capable of recognizing and tracking vocal objects, providing a viable solution for audio-visual fusion target tracking. The mIoU score is 41.5.

However, certain limitations remain and require further improvement in subsequent work. First, in the mask-generation approach, the mask-scoring mechanism can be optimized to enhance object segmentation accuracy. Specifically, a feedback-integrated mask-scoring mechanism could be designed to automatically determine the optimal parameter combination. Second, in AG-STMNet, the learning and representation capabilities of the audio feature extraction network can be further enhanced. Moreover, the fusion level between the audio information and mask decoder can be improved to strengthen the model's ability to suppress environmental noise and interfering sound sources.

Author Contributions: Data curation, Y.Z. (Yunpeng Zuo) and Y.Z. (Yunwei Zhang); investigation, Y.Z. (Yunpeng Zuo) and Y.Z. (Yunwei Zhang); methodology, Y.Z. (Yunpeng Zuo) and Y.Z. (Yunwei Zhang); software, Y.Z. (Yunpeng Zuo); visualization, Y.Z. (Yunpeng Zuo); validation, Y.Z. (Yunpeng Zuo) and Y.Z. (Yunwei Zhang); writing—original draft preparation, Y.Z. (Yunpeng Zuo); writing—review, Y.Z. (Yunpeng Zuo). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key Laboratory of Industrial Intelligence and System of Yunnan Province.

Institutional Review Board Statement: Ethical review and approval were waived for this study for the following reasons. The research involved a non-invasive gait extraction study of quadruped animals based on computer vision. All of the data we used were from publicly available datasets and videos taken in zoos, without contact with the animals. The study did not cause any harm to any animals. According to the type of procedure used, no formal ethical approval was required.

Informed Consent Statement: Not applicable.

Data Availability Statement: AVSbench dataset at <https://github.com/OpenNLPLab/AVSbench> (accessed on 10 July 2022). Ref-AVSbench dataset at <https://gewu-lab.github.io/Ref-AVS> (accessed on 1 July 2024).

Acknowledgments: The authors greatly acknowledge the financial support by the Key Laboratory of Industrial Intelligence and System of Yunnan Province. We would also like to thank the editors for their hard work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhou, J.; Wang, J.; Zhang, J.; Sun, W.; Zhang, J.; Birchfield, S.; Guo, D.; Kong, L.; Wang, M.; Zhong, Y. Audio-visual segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 386–403.
2. Zhang, S.; Zhang, Y.; Liao, Y.; Pang, K.; Wan, Z.; Zhou, S. Polyphonic sound event localization and detection based on Multiple Attention Fusion ResNet. *Math. Biosci. Eng.* **2024**, *21*, 2004–2023. [[CrossRef](#)] [[PubMed](#)]
3. Liu, J.; Liu, Y.; Zhang, F.; Ju, C.; Zhang, Y.; Wang, Y. Audio-visual segmentation via unlabeled frame exploitation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 26328–26339.
4. Chen, T.; Tan, Z.; Gong, T.; Chu, Q.; Wu, Y.; Liu, B.; Lu, L.; Ye, J.; Yu, N. Bootstrapping audio-visual segmentation by strengthening audio cues. *arXiv* **2024**, arXiv:2402.02327. [[CrossRef](#)]
5. Guo, C.; Huang, H.; Zhou, Y. Enhance audio-visual segmentation with hierarchical encoder and audio guidance. *Neurocomputing* **2024**, *594*, 127885. [[CrossRef](#)]
6. Hao, D.; Mao, Y.; He, B.; Han, X.; Dai, Y.; Zhong, Y. Improving audio-visual segmentation with bidirectional generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 2067–2075.
7. Yang, Q.; Nie, X.; Li, T.; Gao, P.; Guo, Y.; Zhen, C.; Yan, P.; Xiang, S. Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 27134–27143.
8. Mao, Y.; Zhang, J.; Xiang, M.; Zhong, Y.; Dai, Y. Multimodal variational auto-encoder based audio-visual segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 954–965.
9. Liu, C.; Li, P.P.; Qi, X.; Zhang, H.; Li, L.; Wang, D.; Yu, X. Audio-visual segmentation by exploring cross-modal mutual semantics. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 7590–7598.
10. Wang, Y.; Sun, P.; Li, Y.; Zhang, H.; Hu, D. Can Textual Semantics Mitigate Sounding Object Segmentation Preference? In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2024; pp. 340–356.
11. Wang, Y.; Zhu, J.; Dong, F.; Zhu, S. Progressive Confident Masking Attention Network for Audio-Visual Segmentation. *arXiv* **2024**, arXiv:2406.02345.
12. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 4015–4026.
13. Mo, S.; Tian, Y. AV-SAM: Segment anything model meets audio-visual localization and segmentation. *arXiv* **2023**, arXiv:2305.01836.
14. Huang, S.; Ling, R.; Li, H.; Hui, T.; Tang, Z.; Wei, X.; Han, J.; Liu, S. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. *arXiv* **2024**, arXiv:2408.15876. [[CrossRef](#)]
15. Bhosale, S.; Yang, H.; Kanojia, D.; Deng, J.; Zhu, X. Unsupervised audio-visual segmentation with modality alignment. In Proceedings of the AAAI Conference on Artificial Intelligence, Edmonton, AB, Canada, 10–14 November 2025; Volume 39, pp. 15567–15575.
16. Liu, J.; Wang, Y.; Ju, C.; Ma, C.; Zhang, Y.; Xie, W. Annotation-free audio-visual segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 5604–5614.
17. Nguyen, K.B.; Park, C.J. SAVE: Segment Audio-Visual Easy way using Segment Anything Model. *arXiv* **2024**, arXiv:2407.02004.
18. Lin, Y.B.; Sung, Y.L.; Lei, J.; Bansal, M.; Bertasius, G. Vision transformers are parameter-efficient audio-visual learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 2299–2309.
19. Xu, S.; Wei, S.; Ruan, T.; Liao, L.; Zhao, Y. Each Perform Its Functions: Task Decomposition and Feature Assignment for Audio-Visual Segmentation. *IEEE Trans. Multimed.* **2024**, *26*, 9489–9498. [[CrossRef](#)]
20. Zhao, X.; Ding, W.; An, Y.; Du, Y.; Yu, T.; Li, M.; Tang, M.; Wang, J. Fast segment anything. *arXiv* **2023**, arXiv:2306.12156.
21. Wu, H.H.; Seetharaman, P.; Kumar, K.; Bello, J.P. Wav2clip: Learning robust audio representations from clip. In Proceedings of the ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 7–13 May 2022; pp. 4563–4567.

22. Oh, S.W.; Lee, J.Y.; Xu, N.; Kim, S.J. Video object segmentation using space-time memory networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9226–9235.
23. Yu, J.; Li, H.; Hao, Y.; Wu, J.; Xu, T.; Wang, S.; He, X. How Can Contrastive Pre-training Benefit Audio-Visual Segmentation? A Study from Supervised and Zero-shot Perspectives. In Proceedings of the BMVC, Aberdeen, UK, 20–24 November 2023; pp. 367–374.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.