Stepping Stones: A Progressive Training Strategy for Audio-Visual Semantic Segmentation

Juncheng Ma[1], Peiwen Sun[2], Yaoting Wang[3], and Di Hu[✉3]

[1] University of Chinese Academy of Sciences
majuncheng21@mails.ucas.ac.cn
[2] Beijing University of Posts and Telecommunications
sunpeiwen@bupt.edu.cn
[3] Gaoling School of Artificial Intelligence, Renmin University of China, Beijing
yaoting.wang@outlook.com
dihu@ruc.edu.cn

Abstract. Audio-Visual Segmentation (AVS) aims to achieve pixel-level localization of sound sources in videos, while Audio-Visual Semantic Segmentation (AVSS), as an extension of AVS, further pursues semantic understanding of audio-visual scenes. However, since the AVSS task requires the establishment of audio-visual correspondence and semantic understanding simultaneously, we observe that previous methods have struggled to handle this mashup of objectives in end-to-end training, resulting in insufficient learning and sub-optimization. Therefore, we propose a two-stage training strategy called Stepping Stones, which decomposes the AVSS task into two simple subtasks from localization to semantic understanding, which are fully optimized in each stage to achieve step-by-step global optimization. This training strategy has also proved its generalization and effectiveness on existing methods. To further improve the performance of AVS tasks, we propose a novel framework Adaptive Audio Visual Segmentation, in which we incorporate an adaptive audio query generator and integrate masked attention into the transformer decoder, facilitating the adaptive fusion of visual and audio features. Extensive experiments demonstrate that our methods achieve state-of-the-art results on all three AVS benchmarks. The project homepage can be accessed at https://ss-aavs.github.io/.

Keywords: Audio-Visual Segmentation · Audio-Visual Semantic Segmentation · Multimodality

1 Introduction

In the real world, audio is consistently associated with its sources, enabling humans to locate sound sources visually based on what they hear. In the past few years, this phenomenon has spurred significant research on Audio-Visual Localization (AVL) [1, 10] aiming to predict the location of sound sources within a

✉Corresponding author.

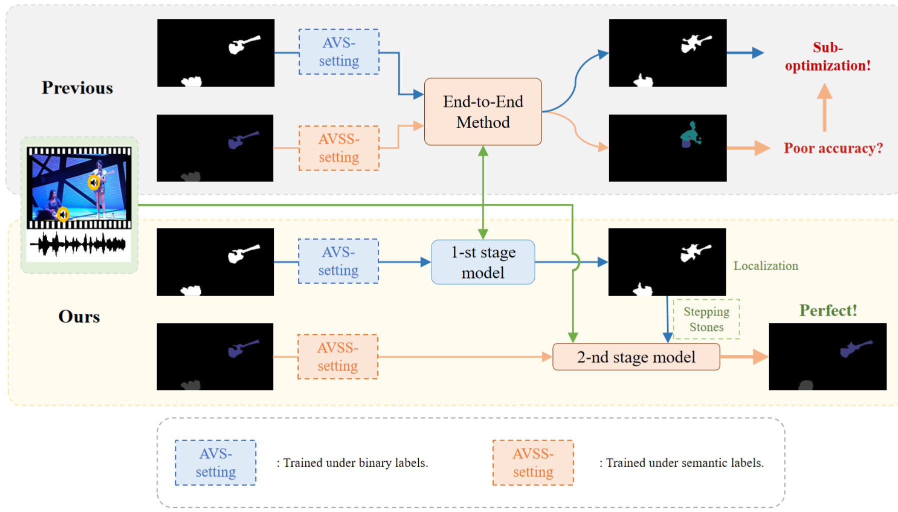

Fig. 1: Comparison between previous methods and our *Stepping Stones* training framework. *High*: Previous end-to-end AVSS approaches result in sub-optimalization on audio-visual alignment. Specifically, when trained under the AVSS-setting, these methods exhibit weaker sound source localization capability compared to training under the AVS-setting. *Low:* our *Stepping Stones* training strategy decomposes the intricate AVSS task into two relatively simple subtasks to be fully learned in the two stages, enhancing the performance significantly.

video, enhancing the human-like perceptual capabilities of machines. However, due to the lack of precise labels, AVL methods tend to capture only a coarse contour of the sound source. Therefore, Audio-Visual Segmentation (AVS) [32] has recently been proposed as a more fine-grained task over AVL, providing pixel-level labels of sound sources. Subsequently, a more challenging task Audio-Visual Semantic Segmentation (AVSS) is proposed by [31] in the pursuit of machine understanding of audiovisual scenes at a semantic level, requiring generating semantic labels for sound sources. In contrast to the traditional Semantic Segmentation (SS) task, AVSS requires achieving audio-visual alignment utilizing the newly introduced audio modality. Compared to AVS, AVSS demands a deeper semantic understanding of the audio-visual scene. The heightened complexity of the objective in AVSS exacerbates the challenge for the model.

AVSBench [32] is the first framework for AVS that employs multi-stage multimodal feature fusion to achieve dense predictions. Subsequent advancements include the integration of audio-queried transformer structures to alleviate receptive field constraints of convolutional approaches [5, 11], and the incorporation of temporal contexts to capture audio-visual spatial-temporal correlations [13]. In addition, [17, 28] use foundation models such as SAM [12] to improve the generalization and performance of the model. However, previous methods may have resulted in suboptimal outcomes and inadequate learning, attributed to the inherent complexity of the AVSS task and the indiscriminate propagation of supervised signals across different modalities during end-to-end training.

Specifically, the AVSS model necessitates semantic-level scene comprehension while simultaneously achieving audio-visual alignment to predict semantic labels of sound sources, effectively combining aspects of AVS and SS. Although the end-to-end approach is widely favored for its convenience, previous end-to-end methods for the AVSS task, while aiming for global optimization, may not sufficiently learn or reach sub-optimization under the supervision of such mixed objectives due to the following two aspects. Firstly, supervised signals are ambiguous during end-to-end optimization. For example, a pixel labeled as *background* may either truly belong to the *background* or represent a silent but potentially semantically significant object such as a *guitar*. Yet, distinguishing between these scenarios during training is not feasible, potentially leading to confusion in semantic understanding when the *background* signal propagates to the entire model in the latter case. On the other hand, the propagation of supervised signals during end-to-end training may be insufficiently utilized. In previous end-to-end approaches, the model needs to simultaneously learn semantic understanding primarily relying on visual cues and audio-guided modal alignment to filter out silence regions, each having its own focal points, potentially resulting in sub-optimization respectively while striving for global optimization. Limitment of end-to-end methods was also corroborated through experiments conducted on AVSBench [32], as depicted in the *high* of Fig. 1, wherein we observed that the model trained under the AVSS task setting exhibits inferior audio-visual alignment (with semantics disregarded) compared to the model trained under the AVS task setting. More results are shown in Fig. 4.

To solve this problem and boost the performance of previous AVSS methods, we propose a simple yet effective two-stage progressive training strategy called **Stepping Stones** (depicted in the *Low* of Fig. 1) to decompose the intricate AVSS task into two relatively simple subtasks to be fully learned in the two stages. In the first stage, the model is trained using binary labels without semantic information, prioritizing the alignment of audio and visual modalities to establish fine-grained correlations at the pixel level. Subsequently, in the second stage, the model is trained with semantic labels as supervision, leveraging the sound source localization results from the first stage as *stepping stones*. These *stepping stones* serve as a shortcut for the model to readily recognize sounding pixels during training, thus focusing on the semantic understanding of the sound source. This approach is generalizable and can be readily applied to existing end-to-end AVSS models. Moreover, acknowledging the considerable error between the first stage results and ground truth during inference, we also introduce the **Robust Audio-aware Key Generator** to robustly integrate auditory information into visual features as keys in the cross-modal cross-attention module. Since the second stage leverages the sound source localization results as prior knowledge, the model can focus solely on the semantic comprehension of a given sound source region, thus overcoming the challenge of ambiguous supervision signals during end-to-end learning. Furthermore, we comprehensively train audio-visual alignment and visually dominant semantic understanding in two separate training phases, which facilitates the optimization of both subtasks of the model, leading to improved global optimization.

In addition, we propose the Adaptive Audio Visual Segmentation model (AAVS), a novel audio-queried transformer built upon Mask2Former [3]. Our framework comprises two key innovations compared with previous methods [5, 11, 13, 16]. Firstly, in previous works, audio queries were either nondiscriminatory or were generated using a module with intricate parameters. We propose an adaptive audio query generator to encode audio queries adaptively, without introducing additional parameters. This approach enhances queries to concentrate on pertinent audio features, thereby improving model performance. Secondly, inspired by [3], we integrate masked attention into the transformer decoder, enabling the model to dynamically adjust its focus on visual features during decoding, which promotes faster convergence and enhanced performance.

To summarise, our main contributions are as follows.

- We propose a simple yet effective progressive training strategy, *Stepping Stones*, for audio-visual semantic segmentation, which can be readily applied to all existing methods. Experiments on several methods demonstrate a notable improvement in validating the strategy's generalization.
- We propose Adaptive Audio Visual Segmentation (AAVS), a novel framework for audio visual segmentation. AAVS incorporates an adaptive audio query generator and integrates masked attention into the transformer decoder, facilitating the adaptive fusion of visual and audio features.
- Extensive experiments show that AAVS significantly outperforms exsiting state-of-the-art approaches in AVSBench-object dataset. Moreover, AAVS trained using our proposed *Stepping Stones* strategy, exhibits considerable performance gains over existing state-of-the-art methods on the more challenging AVSBench-semantic dataset.

2 Related Work

2.1 Audio Visual Localization

Audio-Visual Localization (AVL) aims to predict the location of sound sources within a video, with results typically presented as coarse heatmaps [1, 10]. Prior investigations [23, 26] have employed Audio-Visual Correspondence (AVC) as a self-supervised learning objective, leveraging the inherent alignment between audio and visual features along the temporal dimension. Further advancements by [2, 27] have introduced contrastive learning, utilizing positive and negative instances to enhance sound source localization, emerging as a dominant approach in recent years. To tackle the complexities of multi-source localization, [24] proposed a two-stage training framework incorporating objectives ranging from image-level to class-level granularity. The method facilitates the progressive refinement of multi-source localization, serving as a foundational approach in this field. These pioneering studies have significantly influenced our proposal of a two-stage framework, achieving a step-by-step optimization from localization to semantics for Audio-Visual Semantic Segmentation.

2.2 Audio Visual Segmentation

In contrast to AVL, Audiovisual segmentation represents a more fine-grained task introduced by [32]. AVS offers pixel-level annotations for sound source localization, aiming to achieve a detailed understanding of audiovisual scenes, encompassing both single-source and multi-source scenarios. Furthermore, Audio-Visual Semantic Segmentation emerges as a more challenging task [31], recently proposed as an extension of AVS, with the goal of attaining fine-grained sound source localization and semantic comprehension simultaneously. Existing methods primarily rely on the fusion of visual and audio features. AVSBench [32] is the first framework that employs multi-stage multimodal feature fusion to achieve pixel-level dense predictions. Subsequent advancements [5, 11] have introduced an audio-queried transformer structure to establish global contextual dependencies, thereby mitigating the receptive field limitations inherent in convolutional approaches. Other methods include the incorporation of temporal contexts to capture audio-visual spatial-temporal correlations [13], the utilization of bi-directional generation to establish robust correlations [7], and the integration of generative models to facilitate audio-visual fusion [20]. Additionally, several works [17, 28] leverage the power of large foundation models such as SAM [12] to implement audio-visual segmentation and enhance model generalization. However, previous approaches have overlooked the intricacies inherent in the AVSS task, employing end-to-end models that face challenges in comprehensively mastering both audio visual alignment and semantic comprehension. Consequently, we propose a two-stage strategy, aiming to decompose the AVSS task into two simple subtasks from localization to semantic understanding, thereby facilitating incremental progress toward enhanced model performance.

2.3 Semantic Segmentation

Semantic segmentation is a fundamental task in computer vision, involving the assignment of semantic labels to each pixel in an image. Early methodologies [19, 25] employ an encoder-decoder architecture, which entails downsampling to aggregate global semantic features during encoding and upsampling to recover local detailed features during decoding. Recent advancements such as Mask2Former [3] propose per-mask classification, thereby not only unifying semantic segmentation with other image segmentation tasks but also achieving noteworthy performance. Motivated by these advancements, we propose an Adaptive Audio-Visual Segmentation model for audiovisual segmentation designed to adaptively encode audio queries and dynamically capture visual features using masked attention to attain state-of-the-art performance.

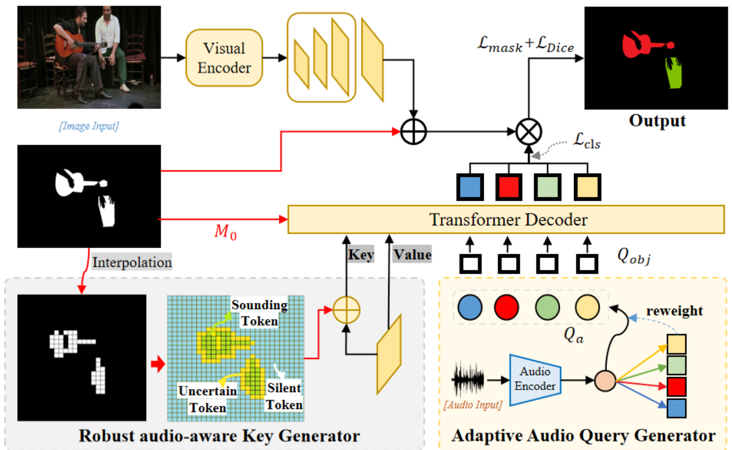

Fig. 2: Overview of AAVS framework. (1) Visual and audio features are extracted by the pre-trained backbone; (2) An Adaptive Audio Query Generator is proposed to generate audio queries; (3) In the transformer decoder, audio-aware queries are integrated with visual feature maps, and masked cross-attention facilitates queries to dynamically adjust the attention range; (4) Finally, refined queries are merged with the mask feature to obtain the final prediction mask. Red rows indicate newly introduced methods to utilize sound source localization results when implementing the Stepping Stones strategy.

3 Methods

In this section, we will first introduce our proposed AAVS framework in Sec. 3.1, including the Adaptive Audio Query Generator, which dynamically fuses audio features with object queries, and the transformer decoder with masked attention to flexibly adjust attention region of the visual feature maps for audio queries. In addition, we will illustrate our training strategy *Stepping Stones* in Sec. 3.2 and show methods to apply it efficiently to the AAVS model in detail.

3.1 Adaptive Audio Visual Segmentation

Feature Extraction. Followed [13, 32], we employ pretrained backbone as visual encoder to extract feature maps across different scales. After that, we utilize the Feature Pyramid Network [14] as a pixel decoder to aggregate feature maps of various resolutions. To integrate the global semantic and the fine-grained features, we iteratively aggregate neighboring resolution feature maps, resulting in merged multi-scale feature $\{F_V^i \in \mathbb{R}^{T\times D\times \frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}}|i\in\{0,1,2,3\}\}$. Here, T, D, H and W denotes the number of frames, embedding dimensions (default to 256) and the original size respectively.

We preprocess the audio input following previous work [28, 32]. Then, we utilize Vggish [8] pretrained on Audioset [6] and a linear projector to extract audio features as $F_A \in \mathbb{R}^{T \times D}$.

Adaptive Audio Query Generator. Most of the existing methods [11,13,16] for generating audio query lack discriminative features in relation to the audio, simply repeating the audio features F_A and combining it with the learnable object query to be fed into a transformer decoder then. A recent work [5] implicitly decomposes the audio features to obtain the audio query using a transformer structure. However, due to the inherent complexity and ambiguity of audio, we believe that this method does not effectively decompose audio components and is parameter redundant.

Therefore, we propose an Adaptive Audio Query Generator without introducing extra parameters to encode adaptive audio queries, which can directly replace the previous *repeat* method. Initially, we define N_q learnable object queries $Q_{obj} \in \mathbb{R}^{N_q \times D}$ and N_q corresponding audio prototypes $P_{audio} \in \mathbb{R}^{N_q \times D}$ to represent N_q potential sound source objects. Each $P^i_{audio}(i \in [1, N_q])$ also serves as the positional embedding for Q^i_{obj} in the transformer decoder enhancing their correlation during training, which also demonstrates that the module introduces no additional parameters.

Subsequently, we dynamically weight F_A based on the cosine similarity with each P^i_{audio} and then integrate the results with Q^i_{obj} to obtain audio-conditioned query $Q^i_a \in \mathbb{R}^D$,

$$Q^i_a = \frac{< P^i_{audio}, F_A >}{|P^i_{audio}||F_A|} F_A + Q^i_{obj}. \tag{1}$$

By this method, the original audio feature F_A can be adaptively fused to the potential sound source query Q^i_{obj} according to corresponding audio prototypes P^i_{audio}, improving the model's sensitivity and discriminative capacity towards audio signals.

Transformer Decoder. In the transformer decoder, we integrate Q_a and multi-scale visual feature $[F^2_V, F^3_V, F^4_V]$ sufficiently to obtain refined Q_{fuse}. Specifically, the decoder consists of four stages, each comprising three transformer layers corresponding to distinct scales of feature maps. Within each layer, we first compute the cross-attention between Q_a and F^i_V using the Masked Multihead Attention [3], which enables the model to focus more on the region to be predicted and thus adaptively adjust the attention range. Subsequently, we perform self-attention on Q_a and feed it into the Feed Forward Network.

Loss. After decoding, we adopt per-mask classification [4] to generate the final segmentation predictions. Specifically, for each Q^i_{fuse}, the mask predictor combines mask feature F^1_V to predict a binary mask, while the class predictor determines the category for this mask. Then, we use the Hungarian algorithm [3] based on the loss function described in Eq. (2) to identify the optimal set of queries that minimizes Eq. (2) and backpropagate the loss

$$\mathcal{L}_{main} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{dice}\mathcal{L}_{dice}. \tag{2}$$

The loss function comprises three components: $\mathcal{L}_{cls}$ represents the cross-entropy loss for mask classification, while $\mathcal{L}_{mask}$ and $\mathcal{L}_{dice}$ denote the cross-entropy loss and Dice loss [22] for mask prediction respectively. We introduce λ_{cls}, λ_{mask}, and λ_{dice} to weight these losses.

Additionally, we employ deep supervision to compute auxiliary losses for the output of each transformer decoder layer to improve performance. The total loss is presented in Eq. (3).

$$\mathcal{L} = \mathcal{L}_{main} + \lambda_{aux}\mathcal{L}_{aux}, \tag{3}$$

Inference. For inference, Q_{fuse} is fed to the mask predictor and class predictor to generate mask prediction $O_{mask} \in \mathbb{R}^{T\times N\times\frac{H}{4}\times\frac{W}{4}}$ and class prediction $O_{class} \in \mathbb{R}^{T\times N\times(C+1)}$, where C donates the total number of classes and 1 donates a extra class *null*. Then, O_{mask} and O_{class} are postprocessed to obtain $O_{pred} \in \mathbb{R}^{T\times(C+1)\times\frac{H}{4}\times\frac{W}{4}}$ followed [3]. Finally, We apply bilinear interpolation to upsample O_{pred} to its original size and take the argmax along the category dimension to obtain the final prediction mask $P \in \mathbb{R}^{T\times H\times W}$.

3.2 Stepping Stones Training Strategy

As mentioned above, the objective of AVSS can be viewed as the combination of AVS and SS, which necessitates the model to simultaneously learn audio-visual correspondence and semantic understanding. However, the integration of these objectives has led in insufficient learning and sub-optimal performance in previous end-to-end methods, prompting us to propose a two-stage training strategy, termed *Stepping Stones*, to tackle the intricate goal as depicted in Fig. 1. In the first stage, we train an AVS model supervised by binary labels to focus on sound source localization. Subsequently, in the second stage, we train an AVSS model supervised by semantic labels, leveraging the first-stage results $\hat{\mathbf{M}}$ as *stepping stones*. Here, the model shifts its emphasis to semantic discrimination of sound source, with audio augmenting visual modal features to enhance performance. By decomposing the AVSS task into two stages of step-by-step learning, we further maximize the capabilities of the end-to-end model and facilitate its comprehensive training. To mitigate the impact of errors, we used ground truth labels $\mathbf{M}$ during training and $\hat{\mathbf{M}}$ exclusively during testing.

Next, we will elaborate on how the second stage model effectively utilizes the first-stage results as *stepping stones* when applying the strategy to AAVS. The modified AVSS model is illustrated in Fig. 2.

Prior Knowledge Input as *Stepping Stones*. In contrast to semantic segmentation, the newly introduced audio modality in the AVSS task plays a crucial role in filtering out silent objects through modal alignment. Consequently, $\mathbf{M}$ offers reliable prior knowledge and mitigates the need for audio-visual alignment within the second-stage model. In the second stage, the model can stand on the *stepping stones* and concentrate further on comprehending visual semantics.

Specifically, we integrate $\mathbf{M}$ with the mask feature F_V^1 in the prediction head. This simple operation serves as a shortcut for the model, aiding in the easy recognition of the sounding region and enhancing its focus on the semantic discrimination of the sound source during training.

Additionally, we utilize $\mathbf{M}$ as the initialization mask for masked attention in the transformer decoder. Previous initialization methods result in notably low accuracy of the original mask, where the initial attention mask is derived from the initial query. An inaccurate initialization mask can hinder the effective confinement of attention to the foreground region in the cross-attention module, potentially leading to slow convergence or model degradation performance. Therefore, we naturally employ $\mathbf{M}$ as prior information to initialize the attention mask, enabling queries to concentrate on region relevant to the sound source.

Robust Audio-aware Key Generator. To mitigate the inevitable error between $\hat{\mathbf{M}}$ obtained from the first stage and ground truth labels $\mathbf{M}$, we devised the Robust Audio-aware Key Generator to fuse audio information into feature maps $[F_V^2, F_V^3, F_V^4]$ as key in cross-modal cross attention to bolster the model's robustness and maximize its utilization of the *stepping stones*.

The module incorporates three learnable embeddings: E_{silent}, $E_{uncertain}$, $E_{sounding}$ and two thresholds τ_1, τ_2 ($\tau_1 < \tau_2$). Initially, $\mathbf{M}$ is resized to match the feature map F_V^i. Subsequently, $\mathbf{M_{resized}}$ is thresholded by τ_1 and τ_2 to produce a index mask. Then, an audio-aware embedding mask $M_{audio} \in \mathbb{R}^{T\times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}} \times D}$ is generated based on the index mask, which is then combined with the F_V^i to derive the audio-aware keys for cross attention. The entire process is illustrated in Eq. (4),

$$M_{audio}^{t,i,j} = \begin{cases} E_{silent} & \text{if } M_{resized}^{t,i,j} < \tau_1 \\ E_{uncertain} & \text{if } \tau_1 \leq M_{resized}^{t,i,j} \leq \tau_2 \\ E_{sounding} & \text{if } M_{resized}^{t,i,j} > \tau_2 \end{cases}, \tag{4}$$

where t represents the current frame, and (i, j) denotes the coordinates of the current pixel. The module enables the query Q_a to dynamically adjust the fusion strategy with visual features based on audio-aware keys during decoding. For instance, in a guitar video scenario, while the original visual key may solely represent the corresponding region as a guitar feature, the audio-aware key can offer supplementary details regarding whether the guitar is *silent*, *uncertain*, or *sounding*, thereby enhancing the model's robustness and capacity.

4 Experiments

4.1 Datasets

AVSBench-Object [32] is an audiovisual dataset for audiovisual segmentation. The dataset consists of two subsets evaluating the S4 and MS3 subtasks respectively, providing pixel-level binary labels about sound sources in video. The S4 subset contains 4,932 videos, with 3,452 videos for training, 740 for validation, and 740 for testing. Conversely, the MS3 subset includes 424 videos, with 286 videos for training, 64 videos for validation, and 64 videos for testing.

AVSBench-Semantic [31] introduces the AVSS subtask as an extension of AVSBench-Object. In addition to providing semantic labels for S4 and MS3 subsets, it also expands the V2 subset. Overall, the AVSBench-Semantic dataset consists of 8,498 videos for training, 1,304 videos for validation, and 1,554 videos for testing. The target objects span 71 categories, encompassing humans, musical instruments, animals, and tools.

Evaluation Metrics. Followed [5,32], we use the Mean Intersection-over-Union (mIoU) and F-score as the evaluation metrics.

Implementation Details. We choose the released Mask2Former [3] model with Swin-B [18] visual backbone pretrained on Ade20k [30] as weight initialization. To effectively leverage the knowledge within the pretrained models, we freeze most of the parameters and introduce adapters [9] for training to achieve faster convergence and improved performance. We utilize the AdamW optimizer with a learning rate of 10^{-4}. The batch size is set to 1. We train the model on the S4 subset for 30 epochs, the MS3 subset for 50 epochs, and the AVSS subset for 30 epochs. To evaluate the effectiveness and generalization of the *Stepping Stones* strategy, we train and test AVSBench and AVSegformer model according to the setting of [5, 32].

4.2 Results and Analysis

We conduct experiments on all three subtasks (S4, MS3, AVSS) of the AVSBench dataset to evaluate the effectiveness of our method and compare it with previous approaches. For fairness, we employ transformer backbone to extract visual features and the AudioSet [6] pre-trained VGGish [8] to extract audio features. The results are presented in Tab. 1.

Quantitative Comparision. We conduct a comprehensive comparison between our AAVS model and existing methods on the AVSBench-Object dataset. The results are presented in Tab. 1. Our AAVS model outperforms previous methods in terms of mIoU and F-score for both the S4 and MS3 subtasks.

Table 1: Quantitative (mIoU, F-score) results on AVSBench dataset with transformer visual backbone. * indicates that the model uses the *Stepping Stones* strategy.

Method	S4		MS3		AVSS		Reference
	mIoU	F-score	mIoU	F-score	mIoU	F-score	
AVSBench [32]	78.74	87.9	54.00	64.5	29.77	35.2	ECCV'2022
AVSC [15]	80.57	88.19	58.22	65.10	-	-	ACM MM'2023
CATR [13]	81.40	89.60	59.00	70.00	32.80	38.50	ACM MM'2023
DiffusionAVS [20]	81.38	90.20	58.18	70.90	-	-	Arxiv'2023
ECMVAE [21]	81.74	90.10	57.84	70.80	-	-	CVPR'2023
AuTR [16]	80.4	89.1	56.2	67.2	-	-	Arxiv'2023
SAMA-AVS [17]	81.53	88.6	63.14	69.1	-	-	WACV'2023
AQFormer [11]	81.60	89.40	61.10	72.10	-	-	IJCAI'2023
AVSegFormer [5]	82.06	89.90	58.36	69.30	36.66	42.00	AAAI'2024
AVSBG [7]	81.71	90.4	55.10	66.8	-	-	AAAI'2024
GAVS [28]	80.06	90.2	63.70	77.4	-	-	AAAI'2024
MUTR [29]	81.5	89.8	65.0	73.0	-	-	AAAI'2024
AAVS (Ours)	**83.18**	**91.33**	**67.30**	**77.63**	**48.50***	**53.20***	ECCV'2024

Specifically, for the S4 subtask, AAVS demonstrates performance improvements of 1.12% in mIoU and 0.93% in the F-score. For the MS3 subtask, AAVS showcases performance improvements of 2.50% in mIoU and 0.23% in the F-score. We attribute these enhancements to the efficacy of our proposed method, which enhances audio-visual modal alignment through adaptive fusion of visual and audio features. It is noteworthy that the improvement on the MS3 subtask is significantly higher than that on the S4 subtask, which we attribute to the greater complexity of the MS3 task scenarios and the adaptive nature of our approach in addressing multi-sound source localization challenges.

For the AVSS subtask, we also conduct experiments to compare our approach with previous work. As depicted in Tab. 1, the AAVS model alone demonstrates performance enhancements of 4.62% mIoU and 4.33% F-score. Moreover, when the *Stepping Stones* training strategy is applied to the AAVS model, performance is further substantially improved by 7.22% mIoU and 6.87% F-score. This significant improvement arises partly from the capabilities of the AAVS model itself and partly from the *Stepping Stones* training strategy, which aids the model in learning audiovisual alignment and semantic understanding more effectively.

Qualitative Comparison Fig. 3 is a comparison of segmentation results obtained from AVSBench and our proposed method. It is evident that our method excels in accurately localizing and segmenting sounding objects, aligning closely with ground truth labels. In the simple S4 task, AAVS effectively identifies the sound source and delineates its boundary with precision. In the more intricate MS3 task, AAVS locates multiple sources well from the audio and provides clear outlines for each. In the most challenging AVSS task, employing the *Stepping*

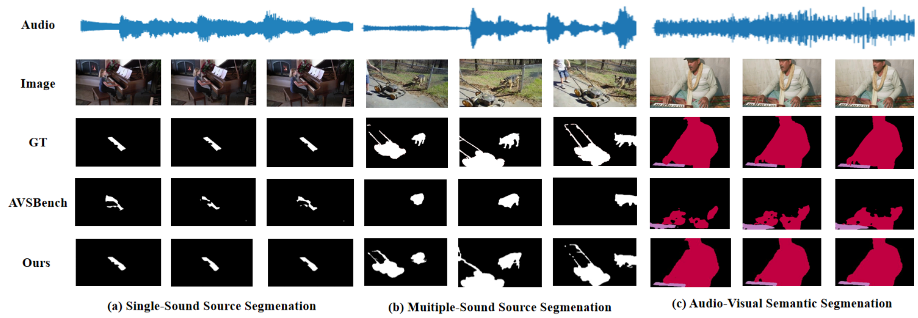

Fig. 3: Qualitative comparison with previous method. The left side is from the S4 subtask, the center is from the MS3 subtask, and the right side is from the AVSS subtask.

Stones training strategy enhances the performance of the AAVS model significantly, yielding segmentation results that closely approximate the ground truth. More qualitative comparison results can be seen in Supplementary Material.

4.3 Generalization of Stepping Stones Training Strategy

To assess the effectiveness and generalizability of the *Stepping Stones* training strategy, we also applied it to the AVSBench and AVSegformer models for the AVSS task with minor modifications. Without introducing any special design, we straightforwardly incorporated the sound source localization results with the mask feature in the prediction head, providing a shortcut for the model. Since the effectiveness of the *Stepping Stones* training strategy relies on the accuracy of the first stage results during inferring, we simulated three levels of accuracy of the first stage results as inputs to the second stage in this section. The results are presented in Tab. 2, revealing that when the accuracy of the first stage results is low, the model's performance evens declines due to the dangerous *stepping stones*. Conversely, when the first stage results demonstrate high accuracy, serving as a reliable stepping stone, the performance of both methods improves. Experiments conducted above only apply the *stepping stones* strategy with minimal modifications to the original method. This underscores the possibility of achieving further performance enhancements by exploring alternative approaches to leveraging the first stage results. The enhancement in performance resulting from the application of the Stepping Stones strategy to AVSBench is also evident in Fig. 4.

4.4 Ablation Study

We conduct ablation experiments to validate the effectiveness of the proposed AAVS model and each key design in the *Stepping Stones* training strategy.

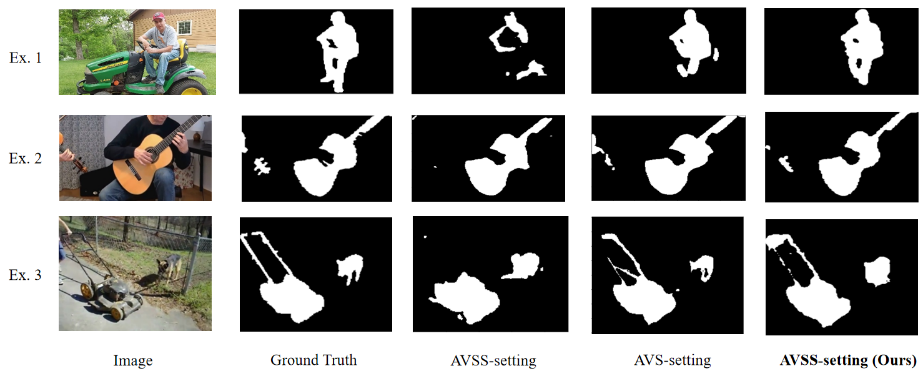

Fig. 4: Previous methods exhibit insufficient learning performance under the AVSS setting. The last column represents the model applying the *Stepping Stones* training strategy. Obviously, the model demonstrates inadequate learning when trained in an end-to-end AVSS setting, whereas the localization accuracy of the semantic mask predicted by the model experiences a notable enhancement following the implementation of *Stepping Stones*.

Table 2: Experiments on previous AVSS Methods. *SS* means *Stepping Stones* strategy in this table. *low*, *high*, and *oracle* corresponded to three levels of mIoU values of the first stage results with *oracle* indicating ground truth labels.

Method	Origin		w/.SS (low)		w/.SS (high)		w/.SS (oracle)	
	mIoU	F-score	mIoU	F-score	mIoU	F-score	mIoU	F-score
AVSBench [32]	29.77	35.2	27.29	29.85	31.48	34.89	36.35	39.01
AVSegformer [5]	36.66	42.0	35.43	39.04	39.44	42.49	46.27	48.04

Ablation of adaptive audio queries. We initially assess the effectiveness of the Adaptive Audio Query Generator within the AAVS model. The experiments were conducted across all three subtasks, with the visual backbone frozen. The results are presented in Tab. 3. It is evident that the inclusion of this module enhances the model's performance compared to using the same audio query across all three subtasks.

Ablation of Stepping Stones training strategy. Finally, we conduct experiments to assess the effectiveness of key components in the *Stepping Stones* training strategy. Since the validity of several components in this section is highly correlated with the accuracy of the AVS results provided from the first stage, our experimental results shed light on the impact of actual pseudo labels and ground truth labels. The pseudo labels, inferred from the trained AAVS model, yield an IoU of 83.18% for S4 labels, 67.50% for MS3 labels, and 72.77% for V2

Table 3: Ablation Experiment on adaptive audio queries.

Method	S4		MS3		AVSS	
	mIoU	F-score	mIoU	F-score	mIoU	F-score
Origin Audio Query	80.71	88.39	65.78	75.21	35.96	40.99
Adaptive Audio Query	**82.51**	**91.33**	**67.50**	**77.63**	**37.13**	**41.72**

Table 4: Ablation Experiment of Stepping Stones training strategy on AAVS. *Actual* refers to sound source localization results inferred from the trained AAVS model, while "Oracle" denotes the utilization of ground truth binary labels. Here, the baseline is the AAVS with audio initialization of the attention mask.

Method	Actual		Oracle	
	mIoU	F-score	mIoU	F-score
final	48.50	53.20	64.85	66.98
w/.o mask fusion	45.23	49.60	54.75	59.65
w/.o robust key	46.44	51.40	60.24	62.28
w/.o robust key+mask fusion	43.43	48.33	51.19	56.19

labels. The results of robustly encoding keys and fusing the sound source localization results with mask features are displayed in Tab. 4. It is evident that all components exhibit enhancements compared to the single-stage AAVS model, highlighting the significant improvement brought about by the *Stepping Stones* training strategy.

5 Conclusion

In this paper, we propose a simple yet effective progressive training strategy called *Stepping Stones*, facilitating a more seamless and comprehensive learning on AVSS task. The method is generalizable and directly applicable to preceding end-to-end methods, thereby enhancing their efficacy in AVSS task performance. In the future, the effectiveness of our training strategy is expected to increase as sound source localization methods with greater accuracy emerge. Additionally, we propose a novel framework AAVS designed for dynamically integrating audio and visual features. Extensive experimental results substantiate the superior performance of AAVS and the *Stepping Stones* training strategy compared to existing state-of-the-art methods.

Discussion of limitations As shown in Sec. 4, there is still a large gap between using sound source localization results obtained from the first stage and ground truth labels. The potential of this training strategy remains ripe for exploration, with future endeavors aimed at delving deeper into enhancing the robustness of the model for *stepping stones* with noise.

References

1. Arandjelovic, R., Zisserman, A.: Objects that sound. CoRR **abs/1712.06651** (2017), http://arxiv.org/abs/1712.06651
2. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: CVPR. pp. 16867–16876 (2021)
3. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1290–1299 (2022)
4. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation **34**, 17864–17875 (2021)
5. Gao, S., Chen, Z., Chen, G., Wang, W., Lu, T.: Avsegformer: Audio-visual segmentation with transformer (2024)
6. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: ICASSP. pp. 776–780. IEEE (2017)
7. Hao, D., Mao, Y., He, B., Han, X., Dai, Y., Zhong, Y.: Improving audio-visual segmentation with bidirectional generation. arXiv preprint arXiv:2308.08288 (2023)
8. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: ICASSP. pp. 131–135. IEEE (2017)
9. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: ICML. pp. 2790–2799. PMLR (2019)
10. Hu, D., Wei, Y., Qian, R., Lin, W., Song, R., Wen, J.R.: Class-aware sounding objects localization via audiovisual correspondence **44**(12), 9844–9859 (2021)
11. Huang, S., Li, H., Wang, Y., Zhu, H., Dai, J., Han, J., Rong, W., Liu, S.: Discovering sounding objects by audio queries for audio visual segmentation. In: IJCAI. pp. 875–883 (2023)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023)
13. Li, K., Yang, Z., Chen, L., Yang, Y., Xiao, J.: Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In: ACM MM. pp. 1485–1494 (2023)
14. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)
15. 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: ACM MM. pp. 7590–7598 (2023)
16. Liu, J., Ju, C., Ma, C., Wang, Y., Wang, Y., Zhang, Y.: Audio-aware query-enhanced transformer for audio-visual segmentation. arXiv preprint arXiv:2307.13236 (2023)
17. Liu, J., Wang, Y., Ju, C., Ma, C., Zhang, Y., Xie, W.: Annotation-free audio-visual segmentation. pp. 5604–5614 (2023)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)

20. Mao, Y., Zhang, J., Xiang, M., Lv, Y., Zhong, Y., Dai, Y.: Contrastive conditional latent diffusion for audio-visual segmentation. arXiv preprint arXiv:2307.16579 (2023)
21. Mao, Y., Zhang, J., Xiang, M., Zhong, Y., Dai, Y.: Multimodal variational auto-encoder based audio-visual segmentation. In: CVPR. pp. 954–965 (2023)
22. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
23. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV. pp. 631–648 (2018)
24. Qian, R., Hu, D., Dinkel, H., Wu, M., Xu, N., Lin, W.: Multiple sound sources localization from coarse to fine. In: ECCV. pp. 292–308. Springer (2020)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
26. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: CVPR. pp. 4358–4366 (2018)
27. Song, Z., Wang, Y., Fan, J., Tan, T., Zhang, Z.: Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In: CVPR. pp. 3222–3231 (2022)
28. Wang, Y., Liu, W., Li, G., Ding, J., Hu, D., Li, X.: Prompting segmentation with sound is generalizable audio-visual source localizer (2024)
29. Yan, S., Zhang, R., Guo, Z., Chen, W., Zhang, W., Li, H., Qiao, Y., He, Z., Gao, P.: Referred by multi-modality: A unified temporal transformer for video object segmentation. AAAI (2023)
30. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR. pp. 633–641 (2017)
31. Zhou, J., Shen, X., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., et al.: Audio-visual segmentation with semantics. arXiv preprint arXiv:2301.13190 (2023)
32. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio–visual segmentation. In: ECCV. pp. 386–403. Springer (2022)