TAViS: Text-bridged Audio-Visual Segmentation with Foundation Models

Ziyang Luo[1] Nian Liu[2,*] Xuguang Yang[1] Salman Khan[2] Rao Muhammad Anwer[2]
Hisham Cholakkal[2] Fahad Shahbaz Khan[2] Junwei Han[1,3]
[1]Northwestern Polytechnical University [2]Mohamed bin Zayed University of Artificial Intelligence
[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

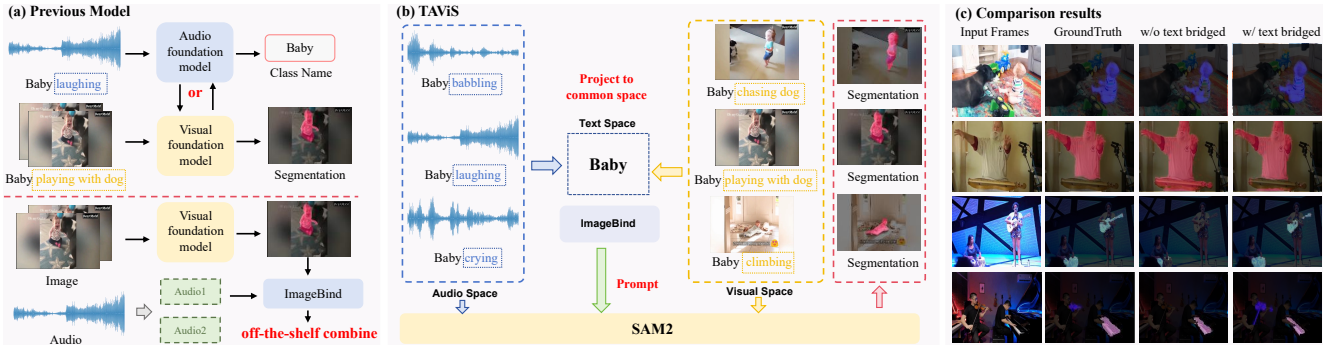

Figure 1. **Illustration of TAViS compared with previous methods.** (1) Previous methods often rely on single-modality foundation models (top) or combine the visual foundation model with ImageBind in an off-the-shelf manner (bottom), which limits their ability to address the misalignment of audio-visual information with significant intra-class diversity. (2) We propose TAViS, a novel framework that leverages two foundation models: ImageBind and SAM2. The text-bridged alignment between visual and audio modalities serves as both prompts and supervision signals for TAViS. (3) We demonstrate the effectiveness of our approach through comparative experiments with and without the text-bridged mechanism.

Abstract

Audio-Visual Segmentation (AVS) faces a fundamental challenge of effectively aligning audio and visual modalities. While recent approaches leverage foundation models to address data scarcity, they often rely on single-modality knowledge or combine foundation models in an off-the-shelf manner, failing to address the cross-modal alignment challenge. In this paper, we present TAViS, a novel framework that ***couples*** *the knowledge of multimodal foundation models (ImageBind) for cross-modal alignment and a segmentation foundation model (SAM2) for precise segmentation. However, effectively combining these models poses two key challenges: the difficulty in transferring the knowledge between SAM2 and ImageBind due to their different feature spaces, and the insufficiency of using only segmentation loss for supervision. To address these challenges, we introduce a text-bridged design with two key components: (1) a text-bridged hybrid prompting mechanism where pseudo text provides class prototype information while retaining modality-specific details from both audio and visual inputs, and (2) an alignment supervision strategy that leverages text as a bridge to align shared semantic concepts within audio-visual modalities. Our approach achieves superior performance on single-source, multi-source, semantic datasets, and excels in zero-shot settings. Source code has been available at https://github.com/Sssssuperior/TAViS*

*Corresponding author: Nian Liu (liunian228@gmail.com)

1. Introduction

Audio-visual joint perception ability is crucial for our understanding of the surrounding environment. Various audio-visual collaboration tasks have emerged in recent years, including audio-visual sound separation [8, 38], visual sound source localization [4, 14, 15, 35], and audio-visual video understanding [16, 19, 23]. Distinct from these tasks, audio-visual segmentation (AVS) [46, 47] aims to generate pixel-wise segmentation maps of sounding objects within a scene, which has garnered increasing attention from the research community.

The primary challenge in AVS lies in effectively align-

ing audio and visual modalities, which requires capturing complex relationships between audio-visual patterns across diverse scenarios. Previous approaches are constrained by limited training data, resulting in incomplete capture of modal relationships. To mitigate this limitation, researchers have introduced large-scale datasets such as AVS-Synthetic [27] and VPO [7]. While pre-training on these datasets has shown promising improvements, it significantly increases computational overhead during training. Thanks to the generalization capabilities of foundation models, several works [25, 27, 33, 44] have utilized image foundation models, such as semantic-SAM [20] or SAM [6], as well as audio foundation models (e.g., Beats [5]), to leverage pre-existing knowledge with minimal additional training on AVS datasets. However, these approaches typically rely on single-modality foundation knowledge, either from audio or image, which hinders effective audio-visual modality alignment. While some previous works [2, 3] attempted to combine segmentation and multimodal foundation models, they employed these models in an off-the-shelf manner for unsupervised settings, where the models operated independently without meaningful interaction, as shown in Figure 1 (a). This reliance on a decoupled combination strategy compromises the transfer of foundation model knowledge.

In this paper, we propose a new model TAViS as shown in Figure 1 (b), designed to couple the knowledge provided by foundation models as a straightforward yet effective solution for the AVS task. However, choosing foundation models for the AVS task is challenging, as it requires both accurate recognition of sound-making objects and detailed segmentation of their boundaries. Our analysis leads us to focus on two complementary foundation models: ImageBind [10] for its robust audio-visual alignment, and SAM2 [36] for its strong generalization capability in complex image and video segmentation. To adapt SAM2 for object-level segmentation from mixed audio guidance, we design an ImageBind-guided query decomposition block (IBQD) that effectively separates audio sources while maintaining ImageBind's well-aligned audio feature space.

After selecting these foundation models, we identify two main challenges in their integration for the AVS task. First, directly transferring ImageBind's alignment knowledge to SAM2's segmentation framework is difficult due to their different feature spaces. Second, using only segmentation loss for supervision is insufficient as it merely implies the need for alignment knowledge. To address the first challenge, we design a text-bridged hybrid prompting technique where text provides class prototype information while audio and visual modalities complement with modality-specific details. For the second challenge of alignment supervision, we propose text-bridged alignment supervision with two separate losses: audio-to-text and image-to-text. While audio and visual modalities contain object-specific information (e.g., different timbre and backgrounds for the same class), text can help extract and align their shared semantic concepts, providing a more effective supervision signal than direct audio-visual alignment. Furthermore, text enables a more generalized classification approach by leveraging similarity comparisons between textual and audio-visual embeddings, rather than relying on a fixed MLP classifier. Based on this similarity-based approach, we further enable TAViS with zero-shot generalization capabilities.

Our main contributions are summarized as follows:

- We **couple** complementary strengths of foundation models for AVS, fusing ImageBind's cross-modal understanding with SAM2's segmentation precision instead of relying on off-the-shelf combined solutions.
- We propose a text-bridged hybrid prompting technique and alignment supervision, transforming audio queries into pseudo-text embeddings while preserving audio-specific cues, and using text as an intermediate bridge for better aligning shared semantic concepts across modalities.
- Our model demonstrates superior performance across diverse AVS benchmarks while excelling in zero-shot settings, establishing a versatile framework that unifies binary, semantic, and zero-shot segmentation capabilities within a single architecture.

2. Related Work

2.1. Audio-Visual Segmentation

In recent years, AVS [45] task has garnered significant attention, as it addresses the challenging problem of locating and predicting pixel-wise maps for sounding objects within a scene. Existing AVS methods can be broadly categorized into three approaches: fusion-based, generative-based, and foundation model-based methods. Fusion-based methods constructed relationships between audio and visual inputs using unidirectional or bidirectional cross-attention mechanisms [9, 21, 22, 24, 44, 45]. For generative-based methods, latent diffusion models [29] or variational auto-encoders [30] were utilized to achieve accurate object segmentation. Foundation model-based methods leveraged visual foundation models such as SAM [18, 27, 33, 34, 37], UniDiffuser [25] and SAM-semantic [20] to obtain enhanced visual knowledge, facilitating the establishment of audio-visual relationships. However, these aforementioned foundation model-based methods primarily focus on visual prior knowledge, leaving a domain gap between audio and visual modalities unaddressed.

In this paper, we propose a novel approach that combines the strengths of Imagebind [10] and SAM2 [36]. Unlike BAVS [25], which utilizes text to align prediction maps, our method designs two separate losses to align text with visual and audio respectively, fully leveraging the text modality for cross-modal alignment. While [2, 3] also combined SAM

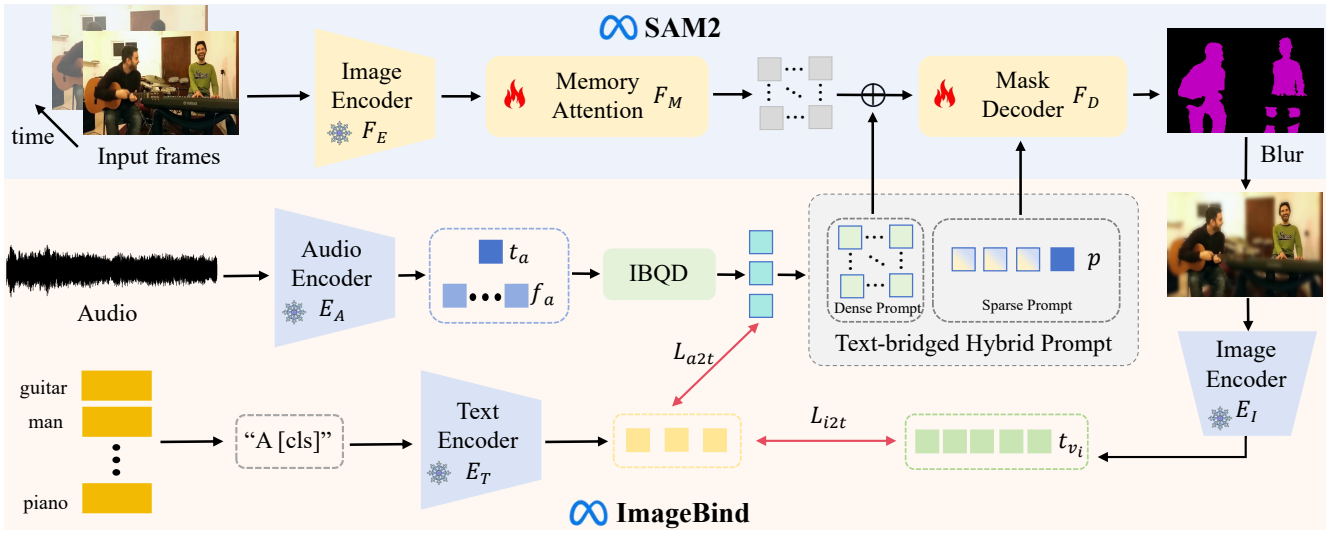

Figure 2. **Overall framework of our proposed model.** Our framework integrates two foundation models: SAM2 for precise segmentation and ImageBind for audio-visual alignment. The input audio is first processed by the audio encoder to extract audio embeddings, which are then decomposed into object-level queries via the IBQD module. Meanwhile, the input image is processed by SAM2. The text-bridged hybrid prompts, including sparse prompt ($\boldsymbol{p}$) and dense prompt ($\boldsymbol{t}_v$), are then generated and fed into the mask decoder to obtain segmentation masks. Finally, two text-bridged alignment losses $\mathcal{L}_{a2t}$ and $\mathcal{L}_{i2t}$ are applied to supervise the audio-to-text and image-to-text relationships, respectively.

and ImageBind for unsupervised audio-visual segmentation, they merely integrated these models in an off-the-shelf manner, which means the two models operate independently without interaction. In contrast, we introduce a text-bridged hybrid prompting design to align ImageBind with SAM2, thereby creating an organically coupled system that leads to more effective optimization.

2.2. Multimodal Joint Representation for Audio

Multimodal joint representation primarily focuses on mapping inputs from different modalities into a shared embedding space. In this section, we specifically focus on audio-related methods. For dual-modality alignment, WavCaps [31] pioneered multimodal representation learning for audio-text pairs, utilizing a large-scale weakly-labeled audio captioning dataset. In the multi-modal context, AudioCLIP [28] and WAV2CLIP [42] extended CLIP's architecture by introducing dedicated audio encoders trained through contrastive learning, leveraging audio-image-text and audio-image pairs respectively. More recent approaches, including ImageBind [10], LanguageBind [48], and OmniBind [41], proposed comprehensive frameworks for integrating multiple modalities by utilizing various data pairs that share common image, language, and text modalities.

Existing AVS methods [25, 27, 33, 44] typically rely on single-modality foundation models for either audio or visual processing independently, resulting in heterogeneous feature spaces that complicate cross-modal alignment. While [32] trained a specialized foundation model for audio-visual tasks, it lacks text alignment capabilities and predicts fixed classes via MLP prediction, which constrains its applicability in real-world scenarios. In contrast to these approaches, our model can perform zero-shot setting, demonstrating superior generalizability for unseen classes through our innovative text-bridged design approach.

3. Methodology

In this work, we propose TAViS, a novel framework that leverages foundation models to achieve precise region segmentation by bridging audio and visual content through text representations. Built upon SAM2 [36] and ImageBind [12] (Section 3.1), our approach introduces three key components: 1) an ImageBind-guided query decomposition module to decompose the mixed audio embedding obtained from ImageBind into object-level queries (Section 3.2). 2) a text-bridged hybrid prompting mechanism that consists of pseudo-text embeddings and audio queries for sparse prompt and image *cls* token for dense prompt (Section 3.3). And 3) two text-bridged alignment loss functions, $\mathcal{L}_{a2t}$ and $\mathcal{L}_{i2t}$, which establish robust audio-visual relationships using text as an intermediate bridge (Section 3.4). The overall architecture of our framework is illustrated in Figure 2.

3.1. Preliminaries

For visual segmentation, we adopt SAM2 [36] as our foundation model. SAM2 consists of three key components: an image encoder F_E to extract visual features, a memory attention module F_M to maintain temporal consistency, and a mask decoder F_D to generate precise segmentation masks for regions of interest. To establish effective audio-visual alignment, we incorporate ImageBind [12] into our frame-

work. Within the ImageBind architecture, we specifically utilize its image encoder E_I, text encoder E_T, and audio encoder E_A. For each modality encoder, we obtain the trunk feature and a *cls* token with a transformer architecture.

Given an audio signal $\boldsymbol{A}$ and its corresponding video frames $\boldsymbol{I} \in \mathbb{R}^{T\times H\times W\times 3}$, where T denotes the total number of frames in the video, we first extract visual features through SAM2's image encoder F_E, generating multi-stage visual features. Concurrently, we process the audio input $\boldsymbol{A}$ through ImageBind's audio encoder E_A to obtain the audio trunk feature $\boldsymbol{f}_a$ and the *cls* token $\boldsymbol{t}_a$. To enhance the correlation between audio and visual modalities, we introduce audio-guided adaptors at each stage of SAM2's visual encoder F_E, following the approach in [27].

3.2. ImageBind-guided Query Decomposition

Previous SAM-based AVS methods [18, 27, 33, 34] typically treat the entire audio feature as a single prompt for the mask decoder. However, the mixing of information from multiple sound sources conflicts with the object-specific queries required by SAM2. A straightforward way to mitigate this issue is to decompose the entire audio feature to object-level prompts with a learnable decomposition module. However, naively doing this can not leverage the well-aligned multimodal knowledge learned by ImageBind. To address this challenge, we propose an ImageBind-guided Query Decomposition (IBQD) to decompose audio features into object-specific queries that preserve the aligned feature space of ImageBind, enabling more precise subsequent audio-visual correspondence.

For the audio trunk feature $\boldsymbol{f}_a \in \mathbb{R}^{F\times C}$ and audio *cls* token $\boldsymbol{t}_a \in \mathbb{R}^{1\times C}$ obtained from audio encoder $\boldsymbol{E}_A$, we introduce learnable queries $\boldsymbol{t}_W \in \mathbb{R}^{N\times C}$ to generate decomposed audio sources. Here, F represents the temporal dimension, C is the channel dimension which represents different frequencies, and N indicates the number of learnable queries. To effectively decompose $\boldsymbol{f}_a$ along the frequency dimension, we employ multi-head cross attention (MCA) mechanisms and utilize weight matrices W_q, W_k, and W_v to generate the query Q, key K, and value V, as follows:

$$\boldsymbol{t}_W = \text{Softmax}((W_q\boldsymbol{t}_W)(W_k\boldsymbol{f}_a^T))(W_v\boldsymbol{f}_a) + \boldsymbol{t}_W. \quad (1)$$

For simplicity, we omit the layer normalization [1] operation here. Subsequently, a linear transformation is applied to $\boldsymbol{t}_W$ to generate object-specific biases, which are added to $\boldsymbol{t}_a$ with boardcast operation and obtain $\boldsymbol{t}'_a \in \mathbb{R}^{N\times C}$. Finally, the $\boldsymbol{t}'_a$ is updated by $\boldsymbol{f}_a$ again with MCA:

$$\begin{aligned} \boldsymbol{t}'_a &= \boldsymbol{t}_a + \text{Linear}(\boldsymbol{t}_W), \\ \boldsymbol{t}'_a &= \text{Softmax}((W_q\boldsymbol{t}'_a)(W_k\boldsymbol{f}_a^T))(W_v\boldsymbol{f}_a) + \boldsymbol{t}'_a. \end{aligned} \quad (2)$$

The decomposed queries $\boldsymbol{t}'_a$ in our IBQD module accurately target individual sound sources, aligning with SAM2's object-centric approach. This two-stage cross-attention mechanism effectively deepens the query decomposition process, allowing for a more contextually-aware separation of sound features. Our decomposed queries are generated as the sum of the original *cls* token $\boldsymbol{t}_a$ and biases, thereby preserving the alignment audio feature space of $\boldsymbol{t}_a$ and preventing potential disruptions in subsequent alignment.

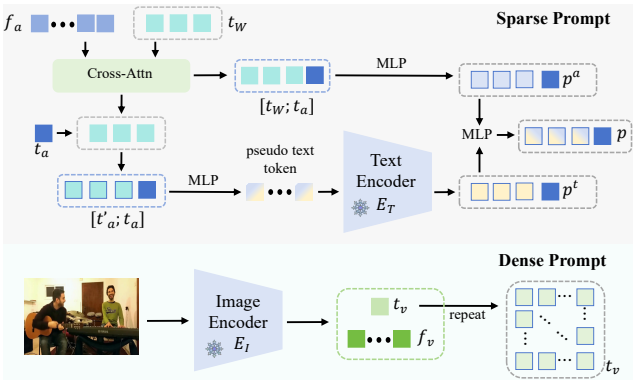

Figure 3. **Illustration of our text-bridged hybrid prompting mechanism.** For the sparse prompt, we combine pseudo text embeddings and specific audio queries generated through MLP. For the dense prompt, we process the image input through ImageBind to obtain the *cls* token, which is repeated and applied across all pixel locations in SAM2's image feature.

3.3. Text-bridged Hybrid Prompting

Prompts play a crucial role in SAM2, providing essential guidance to specify the object of interest to segment. Previous works based on SAM [18, 27, 33, 34] typically process audio inputs through learnable MLPs to generate semantic prompts. However, they may struggle to capture the inherent class prototype nature of audio information required for effective AVS alignment. In AVS scenarios, the initial audio input contains substantial non-semantic information (such as diverse timbres and tones) that can confuse the model. Therefore, we propose novel text-bridged hybrid prompts including sparse prompt and dense prompt, where text provides class prototype information while audio and visual modalities complement with modality-specific details. The overall architecture is shown in Figure 3.

Sparse Prompt. For sparse prompt, we adopt an audio-text dual prompt that leverages text and audio to provide the general object concepts and specific instance details, respectively. We first generate specific object-level text prototype information and scene-level global context following [43]. Specifically, we utilize learnable MLP to the decomposed audio queries $\boldsymbol{t}'_a$ and the audio *cls* token $\boldsymbol{t}_a$. These two transformed tokens are then concatenated and processed by the text encoder E_T to generate the pseudo text prompt $\boldsymbol{p}^t$.

While the pseudo-text prompt captures high-level prototype information, we need to preserve detailed audio characteristics as well. Therefore, we generate specific audio

prompt and global audio context. We concatenate and transform $\boldsymbol{t}_W$ and $\boldsymbol{t}_a$ through another MLP to form the complete audio prompt $\boldsymbol{p}^a$. Finally, we aggregate the $\boldsymbol{p}^t$ and $\boldsymbol{p}^a$ to get the final sparse prompt $\boldsymbol{p}$. The aggregation of these two types of prompts is formulated as follows:

$$\begin{aligned} \boldsymbol{p}^t &= E_T(\text{MLP}[\boldsymbol{t}'_a; \boldsymbol{t}_a]), \\ \boldsymbol{p}^a &= \text{MLP}[\boldsymbol{t}_W; \boldsymbol{t}_a], \\ \boldsymbol{p} &= \text{MLP}([\boldsymbol{p}^a; \boldsymbol{p}^t]), \end{aligned} \tag{3}$$

where [;] denotes concatenation operation, and $\boldsymbol{p}$ is the final spare prompt for SAM2.

To supervise the text embeddings generated from audio, we utilize the trunk feature $E_T(t^t)$ obtained from ImageBind, which helps predict meaningful text information from audio. Here, t^t denotes the ground truth text label with the template "A [cls]". The audio-to-pseudo-text loss $\mathcal{L}_{a2pt}$ is computed as:

$$\mathcal{L}_{a2pt} = \text{MSE}(E_T(\text{MLP}(\boldsymbol{t}'_a)), E_T(t^t)), \tag{4}$$

where MSE denotes the Mean Squared Error loss.

Dense Prompt. Since SAM2's original image feature is not aligned with our audio-text dual prompt, we propose to leverage the image representation from ImageBind as dense prompt to supplement aligned visual knowledge. Specifically, we input the video frame $\boldsymbol{I}$ into ImageBind's image encoder E_I, and extract the *cls* token $\boldsymbol{t}_v$ to represent the global visual information for the whole image. After that, we repeat the image token $\boldsymbol{t}_v$ and add it to each pixel location in the SAM2's image embedding. Please note that the primary role of the global image *cls* token is indeed to enhance high-level alignment between modalities, as ImageBind aligns the final *cls* token, not the feature, here we do not utilize trunk feature as dense prompt.

The sparse prompt effectively transforms the audio information into a text format, which allows us to better capture the core object information while preserving the original audio signal to encode specific semantic characteristics. As for the dense prompt, the visual token provides the contextual information for the frame feature, which benefits the subsequent mask decoding process with the alignment between the image and audio-text modalities. Our text-bridged hybrid prompting strategy leverages the complementary strengths of both the sparse and dense prompts for more effective segmentation.

3.4. Text-bridged Alignment Supervision

While SAM2 and ImageBind demonstrate exceptional capabilities in segmentation and modality alignment respectively, establishing robust audio-visual association between them remains challenging. To leverage these pre-trained models, a straightforward approach would be directly inputting audio and image into the ImageBind encoder for feature alignment. However, this proves suboptimal and hard to optimize as audio contains non-semantic information (e.g., timbre and intonation), while images contain diverse context and appearance information. Both may cause significant intra-class diversity. Moreover, relying solely on segmentation loss for supervision proves insufficient, as it merely implies the need for alignment knowledge without explicitly guiding the model to learn meaningful audio-visual association. Therefore, we propose using text as a "bridge" since it can concisely express high-level prototype information. Our text-bridged alignment supervision consists of two loss terms: $\mathcal{L}_{a2t}$ and $\mathcal{L}_{i2t}$, which establish audio-to-text and image-to-text relationships, respectively.

Audio-to-Text Loss For the audio-to-text loss $\mathcal{L}_{a2t}$, we first format the text input as "A [cls]", where [cls] represents the class labels. Given that the AVS dataset contains P classes (including the background class), we process these P ground truth text inputs t^t through ImageBind's text encoder E_T and projection layer to obtain text embeddings $\text{proj}(E_T(t^t_k))$, where $k \in 0, 1, ..., P$.

Subsequently, the audio queries t'_a are passed through ImageBind's final linear projection layer to obtain their representations $\text{proj}(t'_{a_i})$ in the common embedding space, where $i \in 0, 1, ..., N-1$. Next, we employ the Hungarian matching algorithm to establish optimal assignment between $\text{proj}(E_T(t^t_k))$ and $\text{proj}(t'_{a_i})$. The audio-to-text loss $\mathcal{L}_{a2t}$ is then computed as:

$$\begin{aligned} T_t &= [\text{proj}(E_T(t^t_0)), ...\text{proj}(E_T(t^t_P))], \\ T_a &= [\text{proj}(t'_{a_0}), \text{proj}(t'_{a_1}), ...\text{proj}(t'_{a_{N-1}})], \\ a2t &= T_a \times T_t, \\ \mathcal{L}_{a2t} &= \text{CELoss}(a2t, y), \end{aligned} \tag{5}$$

where CELoss denotes the cross-entropy loss, $a2t \in \mathbb{R}^{N \times P}$, and $y \in \mathbb{R}^N$ represents the index vector of the matched class labels for audio queries.

Image-to-Text Loss Similar to $\mathcal{L}_{a2t}$, we first obtain the text embeddings $\text{proj}(E_T((t^t_k))$ for all class labels. Then, we apply a sigmoid function to the output segmentation map of our model, and use the result to highlight the prediction region on the original image I_i. In this process, we retain the foreground information and apply Gaussian blur to the background, which helps preserve the environmental context without losing too much detail. We then pass the highlighted image I^p_i through ImageBind's image encoder E_I to get t_{v_i} image *cls* token for each object i and utilize projection layer to extract the corresponding visual token $\text{proj}(t_{v_i})$. The image-to-text loss $\mathcal{L}_{i2t}$ can be computed in a similar way:

$$\begin{aligned} T_v &= [\text{proj}(t_{v_0}), \text{proj}(t_{v_1}), ...\text{proj}(t_{v_{N-1}})], \\ i2t &= T_v \times T_t, \\ \mathcal{L}_{i2t} &= \text{CELoss}(i2t, y). \end{aligned} \tag{6}$$

Settings	S4		MS3	
	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$
w/o IBQD	79.6	0.878	58.5	0.664
w/o TbAS	83.6	0.902	64.9	0.716
w/o TbHP	83.9	0.905	65.1	0.713
TAViS	84.8	0.912	68.2	0.759

Table 1. **Ablation studies of key designs of our TAViS.** Key components of our model are progressively removed to evaluate their individual impact. "TbHP" denotes the text-bridged hybrid prompting and "TbAS" represents the text-bridged alignment supervision. All ablation studies are conducted with **224×224** image size.

Here, $i2t \in \mathbb{R}^{N\times P}$ represents the similarity scores between the visual features and the text embeddings, and $y \in \mathbb{R}^N$ denotes the corresponding index vector of the matched labels.

3.5. Overall Training Loss

The final mask loss in our model is divided into two parts. For the object-level loss $\mathcal{L}_{sep}$, we calculate the corresponding prediction mask and compare it with the ground truth using Binary Cross-Entropy loss and IoU loss over N masks. For the binary loss across the entire image $\mathcal{L}_{binary}$, we compute a similar loss applied to a single total binary mask for each image. The overall training loss is expressed as follows:

$$\mathcal{L} = 0.3\mathcal{L}_{a2pt} + 0.5\mathcal{L}_{a2t} + 0.5\mathcal{L}_{i2t} + \sum_{i=0}^{N}\mathcal{L}_{sep} + \mathcal{L}_{binary}. \tag{7}$$

4. Experiment

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate our model on two AVSBench datasets: **AVSBench-object** [47] and **AVSBench-semantic** [46]. The AVSBench-object dataset comprises two scenarios: single sound source segmentation (S4) with 4,932 videos and multiple sound source segmentation (MS3) with 424 videos. The AVSBench-semantic dataset introduces two key differences compared with AVSBench-object. First, it adds a new scenario (V2) containing 6,000 videos with 10 frames per audio clip. Second, it enhances the S4 and MS3 subsets with object-level class label, creating new datasets V1S and V1M, respectively.

Metrics. We adopt two segmentation metrics: $\mathcal{M}_\mathcal{J}$ which measures the overlap between ground-truth masks and predicted ones, and $\mathcal{M}_\mathcal{F}$ which considers both precision and recall in the evaluation. Following previous work [21, 26, 44, 47], we use $\mathcal{M}_\mathcal{J}$ and $\mathcal{M}_\mathcal{F}$ to denote the mean metrics across the entire dataset. On S4 and MS3, we evaluate binary segmentation performance while on AVSS, we report results based on semantic-level metric $\mathcal{M}^I_\mathcal{J}$.

4.2. Implementation Details

We utilize the ViT-L backbone for SAM2 and freeze the parameters of ImageBind and the SAM2 image encoder, while fine-tuning the memory attention and mask decoder. All input frames are resized to 224×224 or 1024×1024. Training is performed using the Adam optimizer [17], starting with an initial learning rate of 1e-4 and a cosine decay scheduler. The batch sizes used are 4 for the MS3 and AVSS datasets, and 10 for the S4 dataset. Training lasts for 80 epochs on both the MS3 and S4 datasets, and 40 epochs on the AVSS dataset.

Settings	S4		MS3	
	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$	$\mathcal{M}_\mathcal{J}$	$\mathcal{M}_\mathcal{F}$
w/o $\mathcal{L}_{a2t}$	83.8	0.905	65.2	0.725
w/o $\mathcal{L}_{i2t}$	84.0	0.904	64.4	0.700
$\mathcal{L}_{a2t}$ +$\mathcal{L}_{i2t}$ + $\mathcal{L}_{a2i}$	83.6	0.903	66.5	0.744
$\mathcal{L}_{a2i}$	83.1	0.895	66.5	0.733
$\mathcal{L}_{a2t}$ +$\mathcal{L}_{i2t}$	84.8	0.912	68.2	0.759

Table 2. **Ablation studies of our TAViS for text-bridged alignment supervision design (TbAS).** The result of the final line represents the final design of our model, and thus yields the same outcome as the final TAViS result.

4.3. Ablation Study

Architecture Design. To demonstrate the efficacy of various components in our model, we conduct ablation studies and report the results in Table 1. We progressively remove key components from our model to analyze their contribution. First, removing the IBQD module leads to significant performance degradation across all scenarios, demonstrating the importance of decomposing audio queries for precise object-level audio-visual alignment. Subsequently, removing the text-bridged alignment supervision (TbAS) results in notable performance drops, highlighting its effectiveness for accurate object localization. Finally, the elimination of the text-bridged hybrid prompting (TbHP) also leads to performance decline, validating the benefits of our design.

Text-bridged Alignment Supervision. As shown in Table 2, we conduct ablation studies on our text-bridged alignment supervision, i.e., $\mathcal{L}_{a2t}$ and $\mathcal{L}_{i2t}$ across two datasets. Removing $\mathcal{L}_{a2t}$ and $\mathcal{L}_{i2t}$ separately leads to performance degradation, highlighting the importance of both audio-to-text and visual-to-text alignments. To further investigate the role of text embedding, we introduce an additional audio-visual alignment loss $\mathcal{L}_{a2i}$, which aligns the proj(t'_{a_i}) and proj(t_{v_i}) on top of our baseline design. However, incorporating this loss harms performance. We hypothesize that this is due to the noisy nature of the audio and visual tokens, which are generated from audio queries and segmentation masks without accurate ground truth supervision. Aligning these noisy tokens likely introduces uncertainty, degrading performance. This hypothesis is further supported by experiments where we solely use $\mathcal{L}_{a2i}$, which results in reduced performance.

To further demonstrate the effectiveness of using text as a bridge, we employ t-SNE [39] to visualize the audio

Models	Backbone	Size	S4		MS3		AVSS
			$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	$\mathcal{M}^{I}_{\mathcal{J}}$
AVSBench [45]	PVT-v2	224	78.7	0.879	54.0	0.645	-
CATR [21]	PVT-v2	224	81.4	0.896	59.0	0.700	-
DiffusionAVS [29]	PVT-v2	224	81.4	0.902	58.2	0.709	-
ECMVAE [30]	PVT-v2	224	81.7	0.901	57.8	0.708	-
BAVS [25]	PVT-v2	224	82.0	0.886	58.6	0.655	32.6
AVSegFormer [9]	PVT-v2	224	82.1	0.899	58.4	0.693	36.7
UFE-AVS [26]	PVT-v2	224	83.2	0.904	59.5	0.676	-
QDFormer [22]	Swin-T	224	79.5	0.882	61.9	0.661	-
COMBO [44]	PVT-v2	224	84.7	0.919	59.2	0.712	42.1
SAM [18]	ViT-H	1024	55.1	0.739	54.0	0.638	-
AV-SAM [33]	ViT-H	1024	40.5	0.566	-	-	-
GAVS [40]	ViT-B	1024	80.1	0.902	63.7	0.774	-
SAMA-AVS [27]	ViT-H	1024	83.2	0.901	66.9	0.754	-
TAViS	ViT-L	224	84.8	0.912	68.2	0.759	**44.2**
TAViS	**ViT-L**	**1024**	**87.0**	**0.926**	**71.2**	**0.796**	-

Table 3. **Quantitative comparison of our TAViS with other AVS methods on three benchmark datasets.** '-' indicates the code is not available and that our metric definition differs. As SAM-based methods do not calculate the semantic-level miou, here we do not list the result. Due to computational resource constraints, we do not report results with 1024 image size on the AVSS dataset, which has 10 frames per video. The best and second-best performances under all settings are **bolded** and bolded, respectively..

Settings	S4		MS3	
	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$
w/o $\boldsymbol{p}^t$	83.6	0.903	63.4	0.714
MLP prediction	84.1	0.904	63.6	0.710
w/o $\boldsymbol{t}_v$	83.3	0.902	63.6	0.701
$\boldsymbol{p}^t$+$\boldsymbol{p}^a$+$\boldsymbol{t}_v$	84.8	0.912	68.2	0.759

Table 4. **Ablation studies of our TAViS for text-bridged hybrid prompting (TbHP).**

Settings	Model	MACs	Size	S4		MS3	
				$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$
SAMA-AVS	SAM-H	598G	1024	83.2	0.901	66.9	0.754
TAViS	SAM-H	334G	224	84.3	0.911	67.3	0.743
SAMA-AVS	SAM2-L	221G	224	83.5	0.902	66.3	0.727
TAViS	SAM2-L	255G	224	84.8	0.912	68.2	0.759

Table 5. **A fair comparison with SAM-based SOTA methods.**

class and visual class embeddings under different settings. For clearer visualization, we focus on the S4 dataset results, which contains single-class images. We specifically exclude classes with insufficient samples, as it hard to cluster. As illustrated in Figure 4, utilizing text as an intermediate bridge leads to more compact clusters within each class, effectively mitigating the impact of modality-specific noise. In contrast, when removing either text-related loss term ($\mathcal{L}_{a2t}$ or $\mathcal{L}_{i2t}$), the direct audio-visual alignment results in more dispersed clusters, indicating the challenge of handling noisy representations without textual guidance.

Text-bridged Hybrid Prompting. Moreover, we investigate the influence of our text-bridged hybrid prompting, which consists of two components: sparse prompt and dense prompt. We first evaluate the impact of sparse prompt by removing the text prompt $\boldsymbol{p}^t$ while only maintaining the audio prompt $\boldsymbol{p}^a$. As shown in Table 4, this ablation leads to performance degradation, highlighting the crucial role of text-based prototype information in segmentation tasks.

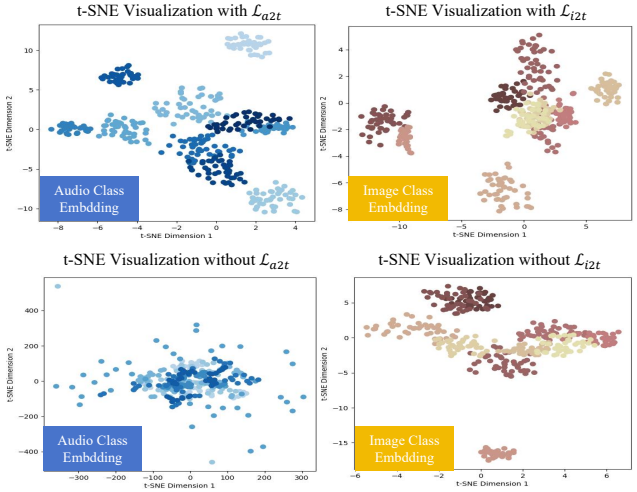

Figure 4. **t-SNE visualization with text-bridge (up) and without text-bridge (down).** Different colors indicate different classes.

Furthermore, we experiment with an alternative text generation approach by removing the text encoder and directly predicting pseudo-text embedding via MLP. This modification results in a significant performance drop on MS3. We conjecture this is because text prediction is highly overfitted on small-scale datasets without the pre-trained knowledge in the text encoder. Similarly, removing the dense prompt $\boldsymbol{t}_v$ also degrades performance, particularly for $\mathcal{M}_{\mathcal{F}}$, demonstrating its vital role in capturing fine-grained segmentation details.

4.4. Comparison with State-of-the-Art Methods

We report the performance comparison of our methods against top-performing state-of-the-art methods, including 14 AVS methods [9, 13, 18, 21, 22, 25–27, 29, 30, 33, 40, 44, 45]. As shown in Table 3, we first compare performance on

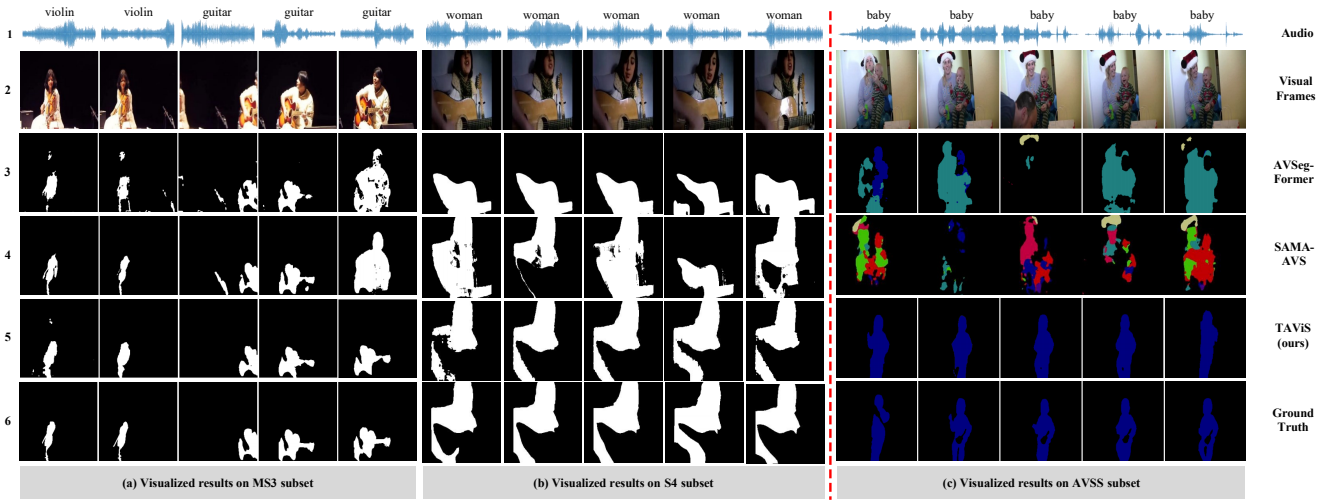

Figure 5. **Qualitative comparison of our model against state-of-the-art AVSS methods.** From top to bottom: (1) audio visibility graphs; (2) visual frames; (3) prediction maps from AVSegFormer [9]; (4) prediction maps from SAMA-AVS [27] (left) and AVSBench [45] (right); (5) prediction maps from our model; and (6) ground truth.

the S4 and MS3 datasets, and our model demonstrates consistent improvements over existing methods on both datasets. Furthermore, we evaluate our model on the AVSS dataset by computing the similarity between audio embedding T_a and text embedding T^t for class prediction of segmented regions. Our approach not only surpasses existing methods on AVSS but also extends SAM-based methods with class prediction capabilities. Notably, increasing the image size from 224 to 1024 significantly improves performance, highlighting the importance of image resolution in SAM(SAM2)-based models, as 1024 matches the foundation model's original training resolution. Figure 5 provides visual comparisons among the top-performing models.

To ensure a fair comparison with previous SAM-based methods, we change our SAM2 foundation model to SAM, as shown in Table 5. Even though with small image size, our model still outperforms the SAM-based SAMA-AVS, demonstrating the effectiveness of our design. Furthermore, we replaced the SAM-based model with SAM2 and retrained the model following the procedure described in the paper. With SAM2, SAMA-AVS showed only a slight improvement, while still performing below our model. These comparative experiments demonstrate that our superior performance is not solely attributed to the use of SAM2, but also stems from our novel model design, which provides greater advantages than the additional pre-training dataset introduced in SAMA-AVS. As our SAM2 version achieves better performance with lower computation cost, we ultimately select SAM2 as our segmentation foundation model.

4.5. Analysis of Zero-shot Generalization Ability

Previous AVSS methods typically predict classes by applying an MLP operation on a fixed set of classes [32, 44], which limits their ability to generalize in real-world scenarios. In contrast, our model leverages ImageBind, which was trained on a broader set of datasets that extend beyond single audio-visual datasets. This enables superior generalization to unseen classes. Following the approach and setting in [11], we train the model on seen classes and evaluate it on unseen classes. To predict the specific class, we compute the similarity of text embeddings with audio and image embeddings respectively, combine them with an addition operation (coefficient set to 1) to obtain the final class prediction. As shown in Table 6, our model outperforms existing OV-AVSS methods with less trainable parameters on zero-shot setting, highlighting the effectiveness of aligning foundation models. While more specialized AVS foundation models may be better suited for the AVSS task [32], the lack of text alignment in such models hinders their generalization ability.

Settings	**Zero-shot $\mathcal{M}^I_{\mathcal{J}}$**	**Trainable Param**
OV-AVSS [11]	22.20	183.6M
TAViS	28.21	54.9M

Table 6. **Comparison with one zero-shot audio-visual segmentation method on AVSBench-OV dataset.**

5. Conclusion

In this paper, we propose TAViS, a novel framework that effectively integrates ImageBind and SAM2 to address the AVS task. First, we design IBQD module that effectively separates audio sources to adapt SAM2 for object-level segmentation. To bridge the two foundation models, we develop a text-bridged hybrid prompting technique and introduce text-bridged alignment supervision through audio-to-text and image-to-text losses. Our model demonstrates superior performance across diverse AVS benchmarks while excelling in zero-shot settings.

Acknowledgments

This work was supported in part by the Noncommunicable Chronic Diseases-National Science and Technology Major Project (Grant 2023ZD0500903, 2023ZD0500900), Key-Area Research and Development Program of Guangdong Province under Grant 2021B0101200001, the National Natural Science Foundation of China under Grant 62136007, 62036011, U20B2065, 6202781, 62036005.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[2] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, Jiangkang Deng, and Xiatian Zhu. Unsupervised audio-visual segmentation with modality alignment. *arXiv preprint arXiv:2403.14203*, 2024. 2

[3] Swapnil Bhosale, Haosen Yang, Diptesh Kanojia, and Xiatian Zhu. Leveraging foundation models for unsupervised audio-visual segmentation. *arXiv preprint arXiv:2309.06728*, 2023. 2

[4] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *CVPR*, pages 16867–16876, 2021. 1

[5] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. *arXiv preprint arXiv:2212.09058*, 2022. 2

[6] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *ICCV workshop*, pages 3367–3375, 2023. 2

[7] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. Unraveling instance associations: A closer look for audio-visual segmentation. In *CVPR*, pages 26497–26507, 2024. 2

[8] Ruohan Gao and Kristen Grauman. Visualvoice: Audio-visual speech separation with cross-modal consistency. In *CVPR*, pages 15490–15500. IEEE, 2021. 1

[9] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *AAAI*, volume 38, pages 12155–12163, 2024. 2, 7, 8

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023. 2, 3

[11] Ruohao Guo, Liao Qu, Dantong Niu, Yanyu Qi, Wenzhen Yue, Ji Shi, Bowei Xing, and Xianghua Ying. Open-vocabulary audio-visual semantic segmentation. In *ACMM*, pages 7533–7541, 2024. 8

[12] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023. 3

[13] Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. In *AAAI*, volume 38, pages 2067–2075, 2024. 7

[14] Di Hu, Rui Qian, Minyue Jiang, Xiao Tan, Shilei Wen, Errui Ding, Weiyao Lin, and Dejing Dou. Discriminative sounding objects localization via self-supervised audiovisual matching. *NeurIPS*, 33:10077–10087, 2020. 1

[15] Di Hu, Yake Wei, Rui Qian, Weiyao Lin, Ruihua Song, and Ji-Rong Wen. Class-aware sounding objects localization via audiovisual correspondence. *TPAMI*, 44(12):9844–9859, 2021. 1

[16] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *ICCV*, pages 5492–5501, 2019. 1

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 2, 4, 7

[19] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weakly-supervised action localization. In *ICLR*, 2020. 1

[20] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023. 2

[21] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *ACMMM*, pages 1485–1494, 2023. 2, 6, 7

[22] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiulian Peng, Rita Singh, Yan Lu, and Bhiksha Raj. Qdformer: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. In *CVPR*, pages 3402–3413, 2024. 2, 7

[23] Yan-Bo Lin, Yu-Jhe Li, and Yu-Chiang Frank Wang. Dual-modality seq2seq network for audio-visual event localization. In *ICASSP*, pages 2002–2006. IEEE, 2019. 1

[24] Yuhang Ling, Yuxi Li, Zhenye Gan, Jiangning Zhang, Mingmin Chi, and Yabiao Wang. Transavs: End-to-end audio-visual segmentation with transformer. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7845–7849. IEEE, 2024. 2

[25] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge. *TMM*, 2024. 2, 3, 7

[26] Jinxiang Liu, Yikun Liu, Fei Zhang, Chen Ju, Ya Zhang, and Yanfeng Wang. Audio-visual segmentation via unlabeled frame exploitation. In *CVPR*, pages 26328–26339, 2024. 6, 7

[27] Jinxiang Liu, Yu Wang, Chen Ju, Chaofan Ma, Ya Zhang, and Weidi Xie. Annotation-free audio-visual segmentation. In *WACV*, pages 5604–5614, 2024. 2, 3, 4, 7, 8

[28] Tanvir Mahmud and Diana Marculescu. Ave-clip: Audioclip-based multi-window temporal transformer for audio visual event localization. In *WACV*, pages 5158–5167, 2023. 3

[29] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Yi-

ran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*, 2023. 2, 7

[30] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *ICCV*, pages 954–965, 2023. 2, 7

[31] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 3

[32] Shentong Mo and Pedro Morgado. Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling. In *CVPR*, pages 27186–27196, 2024. 3, 8

[33] Shentong Mo and Yapeng Tian. Av-sam: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*, 2023. 2, 3, 4, 7

[34] Khanh-Binh Nguyen and Chae Jung Park. Save: Segment audio-visual easy way using segment anything model. *arXiv preprint arXiv:2407.02004*, 2024. 2, 4

[35] Rui Qian, Di Hu, Heinrich Dinkel, Mengyue Wu, Ning Xu, and Weiyao Lin. Multiple sound sources localization from coarse to fine. In *ECCV*, pages 292–308. Springer, 2020. 1

[36] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3

[37] Juhyeong Seon, Woobin Im, Sebin Lee, Jumin Lee, and Sung-Eui Yoon. Extending segment anything model into auditory and temporal dimensions for audio-visual segmentation. *arXiv preprint arXiv:2406.06163*, 2024. 2

[38] Efthymios Tzinis, Scott Wisdom, Tal Remez, and John R Hershey. Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation. In *ECCV*, pages 368–385. Springer, 2022. 1

[39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[40] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. In *AAAI*, volume 38, pages 5669–5677, 2024. 7

[41] Zehan Wang, Ziang Zhang, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang Zhao, and Zhou Zhao. Omnibind: Large-scale omni multimodal representation via binding spaces. *arXiv preprint arXiv:2407.11895*, 2024. 3

[42] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP*, pages 4563–4567. IEEE, 2022. 3

[43] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, pages 15254–15264, 2023. 4

[44] Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In *CVPR*, pages 27134–27143, 2024. 2, 3, 6, 7, 8

[45] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*, 2023. 2, 7, 8

[46] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *IJCV*, pages 1–21, 2024. 1, 6

[47] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *ECCV*, pages 386–403. Springer, 2022. 1, 6

[48] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 3