

AuralSAM2: Enabling SAM2 Hear Through Pyramid Audio-Visual Feature Prompting

Yuyuan Liu¹ Yuanhong Chen² Chong Wang³ Junlin Han¹ Junde Wu¹
Can Peng¹ Jingkun Chen¹ Yu Tian⁴(✉) Gustavo Carneiro⁵

¹ Department of Engineering Science, University of Oxford ² Australian Institute for Machine Learning, Adelaide University
³ Stanford University ⁴ University of Central Florida ⁵ University of Surrey

Abstract

Segment Anything Model 2 (SAM2) exhibits strong generalisation for promptable segmentation in video clips; however, its integration with the audio modality remains under-explored. Existing approaches either convert audio into visual prompts (e.g., boxes) via foundation models, or inject adapters into the image encoder for audio-visual fusion. Yet both directions fall short in human-in-the-loop scenarios due to limited prompt accuracy and increased inference overhead. In particular, these adapter-based methods often suffer from audio prompt dilution, where the signal gradually weakens as it propagates through the network. In this work, we propose AuralSAM2, which integrates audio into SAM2 while largely preserving its promptable segmentation capability. Its core module, AuralFuser, fuses audio and visual features to generate sparse and dense prompts. Guided by audio and built upon SAM2’s feature pyramid, these prompts propagate auditory cues across visual layers, reinforcing cross-modal influence. To further align modalities, we introduce an audio-guided contrastive loss that emphasises auditory relevance in dominant visual features. Our method achieves notable accuracy gains on public benchmarks with only minimal impact on the interactive efficiency of promptable segmentation. Our code is available at <https://github.com/yyliu01/AuralSAM2>.

1. Introduction

Large vision foundation models have emerged as a key advancement in computer vision [4, 4, 40], offering versatile and transferable visual representations across domains. Among them, the Segment Anything Model (SAM) series [23, 42] pioneered promptable segmentation via a human-in-the-loop interactive paradigm. In particular, SAM2 [42] extends this paradigm to video by propagating human-provided visual prompts (e.g., points, boxes) across frames to segment targets of interest throughout a clip.

However, real-world scenarios often require a deeper understanding beyond visual features alone [54]. Auditory signals, which frequently coexist with video frames, are not

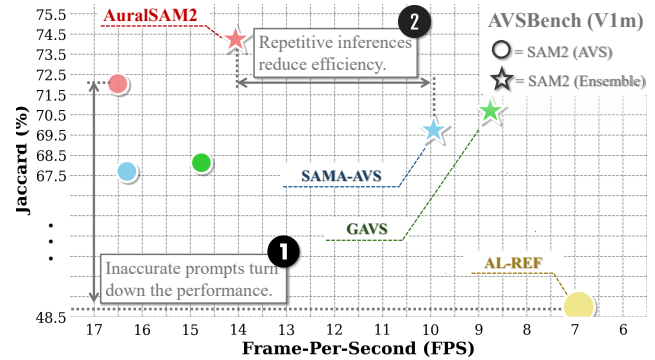


Figure 1. **Prompt Engineering for Integrating Audio Signal in AVSBench (V1m)** [58]. $\circ$ SAM2 (AVS) includes adapter-based methods GAVS [50] and SAMA-AVS [30], along with AL-REF [18], which process audio signals to segment sounding objects. To simulate human-in-the-loop scenarios, $\star$ SAM2 (Ensemble) combines the SAM2 (AVS) results with SAM2 outputs guided by point & box prompts generated from ground truth.

incorporated into SAM2’s inherent design [42]. As a result, users are left to manually scrub through video frames to identify sounding targets, such as a speaking person [2, 47], or an anomalous object making noise [24, 32]. This process is slow [3, 12, 15] and error-prone [49, 53], especially when the object is small [16] or visually ambiguous [45]. In such cases, audio cues serve as a natural guide: they help narrow the search space and stabilise object tracking under occlusion or among look-alike instances. These advantages highlight the potential of audio guidance in promptable segmentation workflows, leading to the core question: *How can we integrate audio guidance into SAM2 without compromising its prompt-driven design for human-AI collaboration?*

A promising direction is Audio-Visual Segmentation (AVS) [58], which explores the semantic relationships between audio and pixel-level visual features in video clips. One common approach [8, 18, 60] is to leverage multimodal foundation models to translate audio into textual descriptions, which are then used to generate visual prompts for SAM2 to localise sounding objects. However, as illustrated in Fig. 1 (1), taken from AL-REF [18], such generated prompts often suffer inaccuracies from hallucination [29].

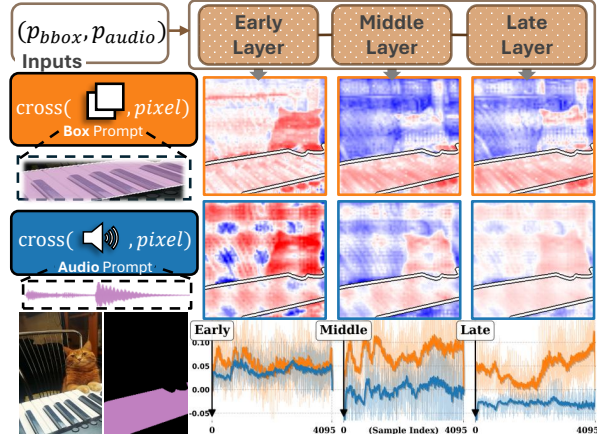


Figure 2. **Audio Prompt Dilution.** The audio prompt signal weakens as it propagates through the SAM2 backbone from [50]. The heatmap visualizes audio–visual cross-attention, and the curve traces its pixel-wise intensity. In contrast, a pretrained bounding box prompt maintains strong alignment throughout the network.

For instance, a box prompt may produce a mask that captures internal patterns instead of the object itself. Moreover, reliance on foundation models increases inference latency and incurs additional costs due to API-based querying [60].

Another line of research [20, 46] introduces audio guidance to SAM2 by injecting adapters into its image encoder, enabling audio–visual feature fusion. However, this integration *alter the intermediate visual features* and degrades SAM2’s promptable segmentation performance. In prompt engineering scenarios with human-in-the-loop, as illustrated in Fig. 1 (⊙), these methods [30, 50] require **repeated SAM2 inferences**: one forward pass to process and fuse audio signals via the adapters (producing audio-conditioned visual features), and another to handle human-provided prompts through SAM2’s promptable interface. This repeated inference significantly slows down the system. For example, ☆ensemble results from [30, 50] are nearly 6.5 FPS slower than their ○AVS results, affecting its real-time feedback performance in practice.

More critically, unlike task-specific methods [10, 13] that tightly couple audio and vision via end-to-end training, adapter-based methods retain a frozen (SAM) backbone and rely on minimal trainable components. This shift poses a unique challenge: audio is not inherently compatible with SAM’s prompt-based design. Comparing with visual prompts, it lacks spatial anchoring and unfolds on a different temporal scale. Simply injecting adapters [30, 50] offers limited control over how audio and pixel signals are fused and propagated across layers. Worse still, the decoder is overwhelmingly dominated by visual features: a single clip yields over 10^6 dense visual tokens, while audio contributes only around 10 coarse embeddings. Taken together, these factors lead to a phenomenon we term audio prompt dilution: as attention propagates deeper into

the model, audio guidance progressively fades. As shown in Fig. 2, while the box prompt maintains strong cross-attention signal with pixel features throughout the decoder, the post-trained audio prompt from [50] weakens progressively, losing its cross-modal correspondence. This is not merely under-utilised audio; it reflects a structural mismatch between how prompts are expected to function in SAM and what audio, in its current form, can reliably deliver in human-in-the-loop workflows.

In this work, we propose *AuralSAM2*, a method designed to enrich SAM2 with audio guidance without compromising its prompt-driven interface. At the core of our method is the AuralFuser module, which is externally attached to the frozen SAM2. This design allows the model to perceive audio signals without modifying image features, thereby avoiding repeated inferences in prompt engineering. To mitigate audio prompt dilution, AuralFuser enhances audio-conditioned attention by generating two complementary sets of feature-level prompts: sparse prompts that capture high-level contextual cues of potential sounding objects, while dense prompts ensure precise pixel-level alignment. These prompts are progressively derived by aligning audio features with a multi-scale feature pyramid built upon patch embeddings from SAM2. This hierarchical design preserves audio guidance throughout the network and strengthens its influence on segmentation. To further counter visual dominance, we introduce an audio-guided contrastive learning (AudioCon) strategy. AudioCon pulls relevant visual features (from pyramid) toward audio prototypes while ignoring visual–visual pairs, reinforcing auditory influence in cross-modal alignment. To summarise, our AuralSAM2’s contributions are:

- We propose AuralFuser, a module that generates audio-conditioned prompts without modifying SAM2’s visual backbone, enabling efficient promptable inference;
- To mitigate audio prompt dilution, AuralFuser constructs sparse and dense prompts through feature pyramid integration, ensuring auditory signal is preserved; and
- We propose AudioCon to further enhance the alignment between audio signals with hierarchical visual features while mitigating the issue of visual dominance.

Our method enables SAM2 to process audio (and optionally language-based audio cues) with minimal efficiency overhead in prompt engineering scenarios. As shown in Fig. 1, AuralSAM2 incurs only a 2.3 FPS drop when adapting visual prompts for the mask decoder, while achieving an Jaccard improvement of 3.9% on AVSBench (V1m) [58], outperforming other SAM2-based SOTA methods.

2. Related Work

Vision Foundation Model methods utilise millions of images and rely on self-supervised learning [4, 40, 48] to enhance feature representation. A notable departure from

this trend is the SAM series [23, 42], which introduces a semi-automated, human-in-the-loop training paradigm. By expanding labeled data through self-generated or human-refined visual prompts (e.g., points and boxes), SAM learns diverse visual patterns across both static images [23] and video clips [42]. In this work, our method is built upon SAM2, chosen for its video-specific design and its strong promptable segmentation capabilities, which we aim to extend to the audio modality without sacrificing human-in-the-loop efficiency.

Audio-Visual Learning (AVL) has been widely studied in deep learning to uncover semantic relationships between audio and visual modalities for enhanced machine perception [61]. It includes tasks such as source separation [7, 33], which extracts distinct sounds from a mixture; binaural audio generation [10], which creates spatial sound from mono or stereo inputs; and sound source localisation [6, 37], which estimates the direction and distance of sound sources. Despite these advances, modeling pixel-level interactions between the two modalities remains a major challenge.

Audio-Visual Segmentation (AVS) has recently been developed to tackle this challenge, with AVSBench [58, 59] serving as the first benchmark, covering both single and multiple sounding sources. The task has since expanded to include zero-shot segmentation for unseen and unheard objects [50], as well as language-aided AVS incorporating textual guidance [51]. *Task-specific AVS models* remain the mainstream approach, with networks retrained from scratch on the AVSBench dataset [58, 59]. Most methods focus on cross-modal fusion, aligning visual features with audio signals before feeding them into a transformer decoder [17, 19, 25, 36], either directly [28, 36] or through learnable audio queries [17, 26]. To further improve alignment, [14] reconstructs audio embeddings from associated visual features, while [26] incorporates temporal cues to enhance spatial correlations between modalities. Contrastive learning [9, 11] has also been explored to strengthen audio-visual associations in the latent space. However, these task-specific AVS models [25, 28, 35] are typically trained on narrow domains, which restricts their generalisability. *AVS for the SAM series* is a promising yet underexplored direction that builds on SAM’s strong generalisation. Existing methods mainly integrate audio via adapters [20], either in the image encoder [38, 46] or across the full architecture [50], enabling fine-tuning on AVS datasets. SAMA-AVS [30] retraining the mask decoder with audio adapters, while GAVS [50] and AV-SAM [38] use audio-visual features as decoder prompts. These approaches modify image features during audio integration, introducing extra inference steps that reduce efficiency. Alternatively, AL-Ref [18] and SAM4AVS [56] use large language or vision-language models [1, 31] to extract audio semantics and generate visual prompts in a zero-shot manner, though they of-

ten suffer from limited accuracy and slow inference. Motivated by these limitations, our proposed AuralFuser integrates audio as an external module without altering the features in the image encoder, thereby avoiding the need for repetitive inference. In addition, our method eliminates reliance on external foundation models by directly generating two sets of feature-level prompts through cross-modal fusion. These prompts effectively guide the SAM2 decoder in capturing sounding objects with both high precision and computational efficiency. Building on this design, Audio-Con further enhances audio-visual alignment by reducing visual dominance impact and reinforcing the guiding role of audio cues via contrastive learning.

3. Method

We define the language-aided AVS dataset [51] as $\mathcal{D} = \left\{ (\mathbf{a}_i, \mathbf{t}_i, \mathbf{v}_i) \mid \mathbf{v}_i = \{(\mathbf{x}_{ij}, \mathbf{y}_{ij})\}_{j=1}^B \right\}_{i=1}^{|\mathcal{D}|}$, where $|\mathcal{D}|$ denotes the number of video clips. The audio signal $\mathbf{a}_i \in \mathcal{A} \subset \mathbb{R}^{N^a \times 2}$ represents a waveform, with N^a being the duration of the audio (based on 16000 Hz sampling rate) with 2 channels. The expression text $\mathbf{t}_i \in \mathcal{T} \subset \mathbb{R}^{1 \times N^t}$ denotes a sentence with N^t words. Each video sequence $\mathbf{v}_i$ consists of B pairs of RGB image $\mathbf{x}_{ij} \in \mathcal{X} \subset \mathbb{R}^{H \times W \times 3}$, with a spatial resolution of $H \times W$, and corresponding pixel-level binarized ground truth masks $\mathbf{y}_{ij} \in \mathcal{Y} \subset [0, 1]^{H \times W}$, representing the sounding object in frame $j \in \{1, \dots, B\}$. Note that in some AVS datasets [58, 59], the language modality $\mathcal{T}$ is unavailable, in which case our work relies solely on audio and visual modalities.

3.1. Preliminaries: SAM2

We define the whole SAM2 as $\mathbf{f}_{\text{SAM2}}^\phi : \mathcal{X} \xrightarrow{\{\mathbf{p}_s, \mathbf{p}_d\}} \mathcal{Y}$, parameterised by ϕ , where $\mathbf{p}_s \in \mathbb{R}^{B \times 5 \times L}$ represents 5 output tokens of dimension L and $\mathbf{p}_d \in \mathbb{R}^{B \times H' \times W' \times L}$ denotes the dense feature maps. Specifically, $\mathbf{p}_s$ comprises 3 mask tokens, 1 object token, and 1 Intersection-Over-Union (IoU) token. Typically, these tokens are concatenated with sparse prompt embeddings (e.g., from points and boxes). The dense features $\mathbf{p}_d$ are computed as the sum of dense (mask) prompt embeddings and visual features, with an output resolution $H' = \frac{H}{16}$ with $W' = \frac{W}{16}$. Since we do not utilise any of the SAM’s prompts in the training, we simplify notation by referring to $\mathbf{p}_s$ as the sparse embeddings and $\mathbf{p}_d$ as the dense embedding in the following discussion. SAM2 is composed of an image encoder represented by $\mathbf{h}_{\text{SAM2}}^{\phi_h} : \mathcal{X} \rightarrow \mathcal{Z}_v$, a memory bank that regularizes the latent feature $\mathcal{Z}_v$, and a mask decoder $\mathbf{g}_{\text{SAM2}}^{\phi_g} : \mathcal{Z}_v \xrightarrow{\{\mathbf{p}_s, \mathbf{p}_d\}} \mathcal{Y}$, such that $\mathbf{f}_{\text{SAM2}}^\phi = \mathbf{h}_{\text{SAM2}}^{\phi_h} \circ \mathbf{g}_{\text{SAM2}}^{\phi_g}$. In the mask decoder $\mathbf{g}_{\text{SAM2}}^{\phi_g}$, two-way cross-attention blocks between $\mathbf{p}_s$ and $\mathbf{p}_d$ occur 3 times, with the sparse and dense features at each block defined as $\mathbf{G} = \{\mathbf{p}_{sk}, \mathbf{p}_{dk} \mid k \in \{1, 2, 3\}\}$. After

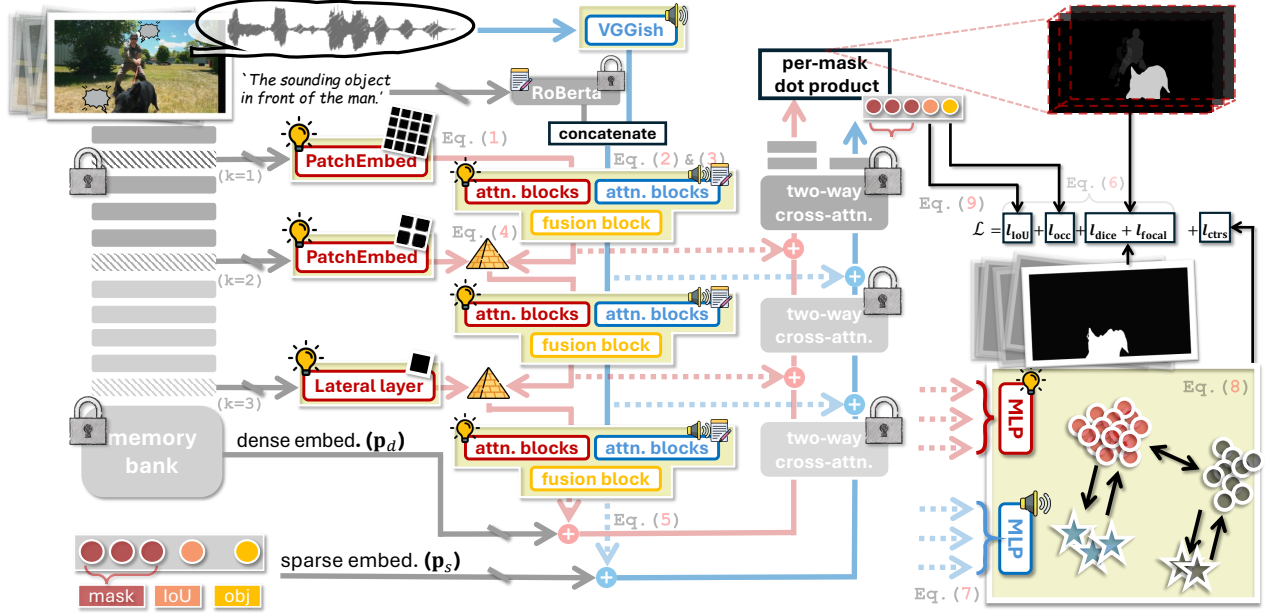


Figure 3. **Illustration of our approach in a language-aided AVS dataset [51].** Audio WAV and text sentences are processed via VGGish [5] and RoBERTa [34], respectively, and then combined. Visual features are extracted from SAM2 in a pyramid structure and processed through PatchEmbedding in Eq. (1) with varying patch sizes (equivalent to the Lateral Layer when $k=3$), then merged using Eq. (4). The visual and audio-text features then undergo self-attention from Eq. (2) and fusion blocks in Eq. (3) to generate sparse and dense feature-level prompts, which guide the mask decoder in capturing potential sounding objects, constrained by the SAM2 loss in Eq. (6) and audio-guided CL (AudioCon) in Eq. (8). Please note that operations based on fused features are highlighted using $\cdots$ and $\cdots$.

processing the final set ($k = 3$) of these tokens through three successive MLPs, the group of predicted binarised masks is computed with the following dot product per mask: $\hat{y}^{\text{mask}} = \mathbf{p}_{d3} \cdot \mathbf{p}_{s3}^{\text{mask}} \in \mathcal{Y}$. The predicted $\hat{y}^{\text{obj}} \in \mathbb{R}$ is a logit derived from $\mathbf{p}_{s3}^{\text{obj}}$ to classify the presence of the target in the current scene. The IoUs of the predicted masks, denoted by $\hat{y}^{\text{IoU}} \in [0, 1]$ are obtained from $\mathbf{p}_{s3}^{\text{IoU}}$ to estimate the overall quality of the output $\hat{y}^{\text{mask}}$.

3.2. AuralFuser

As shown in Fig. 3, AuralFuser processes multi-modal features using pre-trained models as follows:

- 🔊 The *audio waveform* is compressed via $\mathbf{f}_{\text{VGG}}^{\theta^{\text{vgg}}} : \mathcal{A} \rightarrow \mathcal{Z}_a$, where $\mathbf{z}_a \in \mathcal{Z}_a \subset \mathbb{R}^{B \times L}$ and θ^{vgg} denotes the parameter of VGGish [5];
- 📄 The *textual expression* is processed via $\mathbf{f}_{\text{RoBERTa}}^{\psi} : \mathcal{T} \rightarrow \mathcal{Z}_t$, where $\mathbf{z}_t \in \mathcal{Z}_t \subset \mathbb{R}^{N^t \times L}$ and ψ denotes the parameter of RoBERTa [34]; and
- 💡 The *visual features* are extracted after Q-pooling layers [44] to build the pyramid, defined as $\mathbf{Z}_v = \{\mathbf{z}_v^{(k)} \in \mathbb{R}^{B \times \frac{H}{s^{(k)}} \times \frac{W}{s^{(k)}} \times L} \mid \mathbf{s}^{(k)} \in \{4, 8, 16\}, k \in \{1, 2, 3\}\}$, with $\mathbf{Z}_v \subset \mathcal{Z}_v$.

During training, we only update parameters θ (e.g., θ^{vgg} as in [9, 11]), while keeping the text model parameters ψ and SAM2 parameters $\phi = \{\phi^g, \phi^h\}$ fixed. Next, we concatenate the audio and text features to form $\mathbf{z}_c = [\mathbf{z}_a, \mathbf{z}_t]$, where

$\mathbf{z}_c \in \mathbb{R}^{(B+N^t) \times L}$ and apply subsequent operations within our framework that are explained below.

Pyramid Processing: for each $k \in \{1, 2, 3\}$, we process the visual features as follows:

$$\tilde{\mathbf{z}}_v^{(k)} = \mathbf{f}_{\text{PatchEmbed}}^{(k)}(\mathbf{z}_v^{(k)}; \theta_{pe}^{(k)}, p^{(k)}), \quad p^{(k)} \in \{4, 2, 1\}, \quad (1)$$

where $\mathbf{f}_{\text{PatchEmbed}}^{(k)}(\cdot; \theta_{pe}^{(k)}, p^{(k)})$ denotes the patch embedding layer with patch size $(p^{(k)} \times p^{(k)})$ to project all features to the same resolution with $\mathbf{z}_v^{(k)} \in \mathbb{R}^{B \times H' \times W' \times L}$, and it is equivalent to the Lateral Layer when $k=3$ in previous FPN study [27]. Self-attention is then applied independently to both modalities:

$$\begin{aligned} \mathbf{r}_c^{(k)} &= \mathbf{f}_{\text{Attn}^c}^{(k)}(\mathbf{z}_c + \text{Pos}^c; \theta_c^{(k)}), \\ \mathbf{r}_v^{(k)} &= \mathbf{f}_{\text{Attn}^v}^{(k)}(\mathbf{z}_v^{(k)} + \text{Pos}^v; \theta_v^{(k)}), \end{aligned} \quad (2)$$

where $\mathbf{f}_{\text{Attn}^c}^{(k)}(\cdot; \theta_c^{(k)})$ and $\mathbf{f}_{\text{Attn}^v}^{(k)}(\cdot; \theta_v^{(k)})$ are the self-attention blocks for the combined audio-text and visual modalities, respectively, with $\text{Pos}^c \in \mathbb{R}^{(B+N^t) \times L}$ and $\text{Pos}^v \in \mathbb{R}^{B \times H' \times W' \times L}$ denoting their position encodings. Finally, we perform cross-modal fusion as shown below:

$$\mathbf{r}_c^{(k)}, \mathbf{r}_v^{(k)} = \mathbf{f}_{\text{CrossFusion}}^{(k)}(\mathbf{r}_c^{(k)} + \text{Pos}^c, \mathbf{r}_v^{(k)} + \text{Pos}^v; \theta_f^{(k)}), \quad (3)$$

where $\mathbf{f}_{\text{CrossFusion}}^{(k)}(\cdot; \theta_f^{(k)})$ represents the cross-modality fusion block, adapted from TPAVI [58] and the two-way cross-attention fusion mechanism (please see more details

in the Supp. Section 1.3).

For $k \geq 2$, we construct the feature pyramid to integrate early fusion results with late-stage cross-modal alignment, demonstrated as ‘’ in Fig. 3, using:

$$\tilde{\mathbf{z}}_v^{(k)} = \mathbf{f}_{\text{smooth}}^{(k)}(\mathbf{r}_v^{(k-1)} + \tilde{\mathbf{z}}_v^{(k)}; \theta_s^{(k)}), \quad (4)$$

where $\mathbf{f}_{\text{smooth}}^{(k)}(\cdot; \theta_s^{(k)})$ denotes the convolutional smoothing layer with kernel size equal to 1 and is commonly used in the feature pyramid related works [27, 57]. As a result, our approach provides two sets of feature-level prompts. 1) Sparse prompts represent visual-language informed audio features $\mathbf{R}_a = \{\mathbf{r}_a^{(k)} = \text{Select}_a(\mathbf{r}_c^{(k)}) \in \mathbb{R}^{B \times L} \mid k \in \{1, 2, 3\}\}$, where $\text{Select}_a(\cdot)$ is the function that extracts the audio feature $\mathbf{r}_a^{(k)}$ from the combined representation $\mathbf{r}_c^{(k)}$, based on its original position from $\mathbf{z}_a$ in $\mathbf{z}_c$. These features encode global context by capturing the visual data relevant to audio and language modalities. 2) Dense prompts correspond to audio-language enriched visual features $\mathbf{R}_v = \{\mathbf{r}_v^{(k)} \in \mathbb{R}^{B \times H' \times W' \times L} \mid k \in \{1, 2, 3\}\}$, which provides pixel-level identification of all potential sounding objects within the scene.

Hierarchical Prompting. We progressively integrate the prompt sets $\mathbf{r}_a^{(k)}$ and $\mathbf{r}_v^{(k)}$ during the two-way cross-attention blocks in $\mathbf{g}_{\text{SAM2}}^{\phi_g}$ as follows:

$$\begin{aligned} \tilde{\mathbf{p}}_{sk}^{\text{mask}} &= \mathbf{p}_{sk}^{\text{mask}} + \mathbf{r}_a^{(k)}, & \tilde{\mathbf{p}}_{sk} &\in \mathbb{R}^{B \times 5 \times L}, \\ \tilde{\mathbf{p}}_{dk} &= \mathbf{p}_{dk} + \mathbf{r}_v^{(k)}, & \tilde{\mathbf{p}}_{dk} &\in \mathbb{R}^{B \times H' \times W' \times L}, \end{aligned} \quad (5)$$

where $\mathbf{G} = \{(\tilde{\mathbf{p}}_{sk}, \tilde{\mathbf{p}}_{dk}) \mid k \in \{1, 2, 3\}\}$ and we only update the mask token $\mathbf{p}_{sk}^{\text{mask}}$ and $\mathbf{p}_{dk}$ in $\mathbf{g}_{\text{SAM2}}^{\phi_g}$. While the other tokens (i.e., $\mathbf{p}_s^{\text{IoU}}$, $\mathbf{p}_s^{\text{object}}$) can still learn to capture the correct feature via self-attention blocks in $\mathbf{h}_{\text{SAM2}}^{\phi_h}$. As a result, we follow the training pipeline in SAM2 with the loss:

$$\begin{aligned} \ell_{\text{SAM2}}(\mathcal{D}, \theta^{\text{vgg}}, \theta^{(k)}) &= \ell_{\text{focal}}(\hat{\mathbf{y}}^{\text{mask}}, \mathbf{y}) + \ell_{\text{dice}}(\hat{\mathbf{y}}^{\text{mask}}, \mathbf{y}) \\ &+ \ell_{\text{IoU}}(\hat{\mathbf{y}}^{\text{IoU}}, \mathbf{IoU}(\hat{\mathbf{y}}^{\text{mask}}, \mathbf{y})) + \ell_{\text{occ}}(\hat{\mathbf{y}}^{\text{obj}}, \mathbb{I}(\mathbf{y} > 0)), \end{aligned} \quad (6)$$

where $\hat{\mathbf{y}}^{\text{mask}}$, $\hat{\mathbf{y}}^{\text{obj}}$ and $\hat{\mathbf{y}}^{\text{IoU}}$ are defined in the Preliminaries section, $\mathbb{I}(\mathbf{y} > 0) \in \{0, 1\}$ is a binary indicator determining the presence of a foreground object in the label $\mathbf{y}$, and $\mathbf{IoU}$ represents the IoU calculation metric. For further details on this loss, we refer to the SAM2 paper [42].

3.3. Audio-guided CL (AudioCon)

Unlike previous contrastive objectives that treat both modalities symmetrically, AudioCon *privileges* audio as the anchor and only repels visual negatives. This design directly addresses the visual dominance observed in SAM2, ensuring that the most salient clusters in the latent space are organised around audio cues rather than purely visual similarities. In particular, we utilise two MLPs to project the entire

feature sets of $\mathbf{R}_a$ and $\mathbf{R}_v$ into the same embedding space with:

$$\mathbf{e}_a = \mathbf{f}_{\text{proj}^a}(\mathbf{r}_a^{(k)}; \theta_{pa}), \quad \mathbf{e}_v = \mathbf{f}_{\text{proj}^v}(\mathbf{r}_v^{(k)}; \theta_{pv}), \quad (7)$$

where the audio modality embedding $\mathbf{e}_a \in \mathbb{R}^{B \times C}$ contains frame numbers (B) of embedding features, each with dimension C . The visual modality embedding $\mathbf{e}_v \in \mathbb{R}^{B \times H' \times W' \times C}$ has a significantly larger number of embedding features compared to the audio modality, with $B \times H' \times W' \gg B$. Based on the label $\mathbf{y}$, we thus can construct the audio embedding set $\mathcal{E}_a = \{(\mathbf{e}_b^+, \mathbf{y}_b) \mid b = 1, 2, \dots, B\}$; and similarly, we can construct the visual embedding set $\mathcal{E}_v = \{(\mathbf{e}_b^v, \mathbf{y}_b^{(\omega)}) \mid b = 1, 2, \dots, B\}$, where Ω is the lattice of ground truth and ω denotes a pixel-level position with $\omega \in \Omega \subset \mathbb{R}^{H' \times W'}$. Thus, the AudioCon is defined as:

$$\begin{aligned} \ell_{\text{ctrs}}(\mathcal{D}, \theta_{pa}, \theta_{pv}) &= \frac{1}{|\mathcal{E}_v|} \frac{1}{B} \sum_{(\mathbf{e}_b^v, \mathbf{y}_b^{(\omega)}) \in \mathcal{E}_v} \sum_{\substack{(\mathbf{e}^+, \mathbf{y}_b) \in \mathcal{E}_a \\ \mathbb{I}(\mathbf{y}_b = \mathbf{y}_b^{(\omega)})}} \\ &- \log \frac{\exp(\mathbf{e} \cdot \mathbf{e}^+ / \tau)}{\exp(\mathbf{e} \cdot \mathbf{e}^+ / \tau) + \sum_{\substack{(\mathbf{e}^-, \mathbf{y}_b^{(\omega)-}) \in \mathcal{E}_v \\ \mathbb{I}(\mathbf{y}_b^{(\omega)-} \neq \mathbf{y}_b^{(\omega)})}} \exp(\mathbf{e} \cdot \mathbf{e}^- / \tau)}. \end{aligned} \quad (8)$$

where τ is a temperature parameter and $\mathbb{I}(\cdot)$ indicates whether there is a (pixel-level) foreground object matching the current frame’s audio. Unlike previous works [9, 11] that apply InfoNCE [39] to the entire latent space (i.e., $\mathcal{E}_v \cup \mathcal{E}_a$), our AudioCon mitigates modality imbalance by pulling visual embeddings toward relevant audio $\mathbf{e}^+$ while pushing them away from other visual samples $\mathbf{y}_b^{(\omega)-}$. This implementation prevents the model from overemphasizing attraction between pixel-level visual embeddings in $\mathcal{E}_v$. Instead, it aggregates visual features using audio embeddings as central prototypes, thereby ensuring that visual features cluster around meaningful auditory cues. We include t-SNE visualisations in Supp. Section 4.1 to show this effect.

3.4. Training Objective

The training of our AuralSAM2 minimises the following loss function:

$$\mathcal{L}(\mathcal{D}, \theta) = \ell_{\text{SAM2}}(\mathcal{D}, \theta^{\text{vgg}}, \theta^{(k)}) + \ell_{\text{ctrs}}(\mathcal{D}, \theta_{pa}, \theta_{pv}), \quad (9)$$

where $\theta^{(k)} = \{\theta_{pe}^{(k)}, \theta_c^{(k)}, \theta_v^{(k)}, \theta_f^{(k)}, \theta_s^{(k)} \mid \text{if } k \geq 2\} \mid k \in \{1, 2, 3\}$. During the optimisation, we only supervise the mask with the lowest segmentation loss in ℓ_{SAM2} .

4. Experiment

Experimental setup. With language-aided AVS, we evaluate our method on *Ref-AVS* [51] benchmark, which includes 4,002 video clips and 20,261 expressions. Each expression corresponds to a unique object, with 14,117

Table 1. **Comparison with SOTA on the Ref-AVS dataset.** Methods based on SAM [23] are shown in mauve, those based on SAM2 [42] in yellow, while other entries use task-specific models. The † indicates our reimplementation and * denotes methods utilising SAM’s zero-shot capability. The best results are marked in red, and the second best are underlined.

Method	Backbone	Ref-AVS [51]									
		Seen			Unseen			Mix			Null
		$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{J} \& \mathcal{F} \uparrow$	$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{J} \& \mathcal{F} \uparrow$	$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{J} \& \mathcal{F} \uparrow$	$\mathcal{S} \downarrow$
TPAVI [58] [ECCV 2022]	PVT-v2	23.20	51.1	37.2	32.36	54.7	43.5	27.78	52.9	40.3	0.208
AVSegFormer [13] [AAAI 2024]	PVT-v2	33.47	47.0	40.2	36.05	50.1	43.1	34.76	48.6	41.7	0.171
EEMC [51] [ECCV 2024]	Swin-b	34.20	51.3	42.8	49.54	64.8	57.2	41.87	58.1	50.0	0.007
GAVS [50] [AAAI 2024]	ViT-h	28.9	49.8	39.35	29.8	49.7	39.8	29.4	49.8	39.6	0.190
SAMA-AVS [50] [WACV 2024]	ViT-h	39.2	56.2	47.7	47.5	56.6	52.1	43.4	56.4	49.9	0.130
TSAM [41] [CVPR 2025]	ViT-h	43.4	56.8	50.1	54.6	66.4	60.5	49.0	61.6	55.3	0.017
Ours SAM (w/ AuralFuser)	ViT-h	48.26	60.28	54.27	57.91	68.95	63.43	53.09	59.10	58.85	0.053
GroundedSAM2* [43] [arxiv 2024]	Hiera-b+	28.5	39.9	34.2	59.8	68.1	63.9	44.2	54.0	49.1	0.277
GAVS† [50] [AAAI 2024]	Hiera-b+	48.0	54.6	51.3	59.2	65.8	62.5	53.6	60.2	56.9	0.076
SAMA-AVS† [50] [WACV 2024]	Hiera-b+	49.5	56.7	53.1	60.6	66.4	63.5	55.1	61.5	58.3	0.103
SAM2-LOVE [52] [CVPR 2025]	Hiera-l	43.5	51.9	47.7	66.5	72.3	69.4	55.0	62.1	58.5	0.230
Ours AuralSAM2	Hiera-b+	53.16	58.83	56.00	63.45	70.44	66.95	58.31	64.64	61.48	0.129
	Hiera-l	56.16	61.19	58.68	68.69	74.36	71.53	62.43	67.78	65.11	0.065

Table 2. **Comparison with SOTA on the AVSBench dataset.** Methods employing SAM [23] are in mauve, SAM2 [42] in yellow, and the rest are task-specific models. The † denotes our reimplementation, # represents grounding semantic information to the class-agnostic mask via [35], and * denotes methods utilising SAM’s zero-shot capability. The best results are in red and the second best are underlined.

Method	Backbones (audio & visual)	AVSBench [58, 59]							
		V1(single)		V1(multiple)		V2 (binary)		V2 (semantic)	
		$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$
AVSegFormer [13] [AAAI 2024]	VGGish PVT-v2	82.1	89.9	58.4	69.3	-	-	36.7	42.0
AVS-BiGen [14] [AAAI 2024]	VGGish PVT-v2	81.7	90.4	55.1	66.8	64.3	75.9	-	-
Step.-Stones [35] [ECCV 2024]	VGGish Swin-b	83.2	91.3	67.3	77.6	-	-	48.5#	53.2#
SAM4AVS* [56] [BMVC 2023]	VGGish PVT-v2	51.2	61.5	41.8	47.8	-	-	-	-
COMBO* [55] [CVPR 2024]	VGGish PVT-v2	84.7	91.9	59.2	71.2	-	-	42.1	46.1
GAVS [50] [AAAI 2024]	VGGish ViT-b	80.1	90.2	63.7	77.4	67.7	78.8	-	-
SAMA-AVS [30] [WACV 2024]	VGGish ViT-h	81.5	88.6	63.1	69.1	-	-	-	-
Ours SAM (w/ AuralFuser)	VGGish ViT-h	84.78	91.92	65.22	79.13	70.24	81.63	49.52#	54.88#
AL-Ref* [18] [AAAI 2025]	Beats Hiera-l	70.5	81.1	48.6	53.5	59.2	66.2	36.0	39.8
GAVS† [50] [AAAI 2024]	VGGish Hiera-b+	83.64	92.47	68.13	79.07	73.58	84.04	-	-
SAMA-AVS† [30] [WACV 2024]	VGGish Hiera-b+	82.11	90.58	67.70	78.93	74.28	84.35	-	-
Ours AuralSAM2	VGGish Hiera-b+	85.01	92.16	72.04	81.46	76.78	85.38	50.23#	55.16#
	VGGish Hiera-l	86.62	93.34	75.58	84.12	79.09	86.84	50.57#	56.03#

training and 4,770 test cases. The test set is divided into 2,288 seen-object cases for performance evaluation, 1,454 unseen-object cases for generalisation assessment, and 1,028 null cases where the referenced object is absent or not visible. We also evaluate our method on the **AVS-Bench** [58] dataset without language modality, which comprises two subsets: *V1s* and *V1m*, representing single and multiple sounding sources, respectively. The *V1s* subset consists of 3,452 training clips, 740 validation clips, and 740 test clips, while the *V1m* subset includes 296 training cases, 64 validation cases, and 64 test cases, both evaluated in a binary class-agnostic setting. The extended V2 [59] subset builds upon *V1s* and *V1m*, introducing 12,356 video clips across 70 semantic categories.

Metrics. We use the average Jaccard index ($\mathcal{M}_{\mathcal{J}}$) and F-Score ($\mathcal{M}_{\mathcal{F}}$) for evaluating segmentation performance in AVSBench [58], along with an additional Square Root of the Ratio measurement ($\mathcal{S}$) in Ref-AVS [51].

Implementation Details. Our experiments are built upon the SAM2 framework [42] using both the Hiera.base+ and Hiera.large backbones. Following previous SAM-based methods [50], we use an input image resolution of 1024x1024 and a batch size of one across all datasets. Given the limited exploration of SAM2 within AVS, we have re-implemented previous SOTA methods [30, 50] based on their code. During training, the learning rate is set to $1e^{-4}$, with a poly learning rate decay following $(1 - \frac{\text{iter}}{\text{max iter}})^{0.9}$. Consistent with SAM2 [42], we set 20 : 1 : 1 : 1 for the linear combination for ℓ_{focal} , ℓ_{dice} , ℓ_{IoU} and ℓ_{occ} in Eq. (6). For contrastive learning, a three-layer projector is used for both audio and visual features, with an output dimension of 64. The temperature value is set to $\tau = 0.10$ in Eq. (8) and remains constant throughout all experiments. Please refer to Supp. Section 1 for more implementation details and to Supp. Section 3 for results with other backbones.

Table 3. **Ablation Studies** on AVSBench and Ref-AVS using Hiera [44] large backbones. The first row presents results based solely on the visual modality , while the following rows show outcomes from cross-modal fusion with audio  or optional language  modalities. The subsequent two rows illustrate the effect of employing a multi-scale feature pyramid  arranged from bottom to top, with the bottom row further incorporating audio-guided contrastive learning.

Ablations	Pyramid	AVSBench [58, 59]						Ref-AVS [51]					
		V1 (single)			V1 (multiple)			Seen			Unseen		
		$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{J} \& \mathcal{F} \uparrow$	$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{J} \& \mathcal{F} \uparrow$	$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{J} \& \mathcal{F} \uparrow$	$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{J} \& \mathcal{F} \uparrow$
	-	83.41	91.36	87.39	61.50	73.09	67.30	43.77	48.01	45.89	64.33	69.98	67.16
  	-	84.55	92.08	88.32	71.52	79.57	75.55	53.36	57.49	55.43	66.94	72.18	69.56
  	 =2	85.96	92.97	89.47	73.42	81.94	77.68	54.67	58.91	56.79	67.81	72.86	70.34
  	 =3	86.33	93.27	89.80	74.43	82.76	78.60	55.32	60.69	58.00	67.74	73.92	70.83
   AudioCon	 =3	86.62	93.34	89.98	75.58	84.12	79.85	56.16	61.19	58.68	68.69	74.36	71.53

Table 4. **Ablation Studies on CL** in AVSBench [58] dataset based on Hiera.L backbone. Best results are highlighted in red.

Method	AVSBench (V1m)			
	$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{J} \& \mathcal{F} \uparrow$	Δ
w/o CL	74.43	82.76	78.60	-
Ours w/ SupCon	74.86	83.29	79.08	0.48 $\uparrow$
Ours w/ AudioCon	75.58	84.12	79.85	1.25 $\uparrow$

Table 5. **Ablation study of prompts** on AVSBench [58] using the Hiera.L backbone. Best results are shown in red.

Method	AVSBench (V1m)			
	$\mathcal{M}_{\mathcal{J}} \uparrow$	$\mathcal{M}_{\mathcal{F}} \uparrow$	$\mathcal{J} \& \mathcal{F} \uparrow$	Δ
Ours	75.58	84.12	79.85	-
w/o Sparse Prompts	67.06	76.51	71.79	8.06 $\downarrow$
w/o Dense Prompts	63.32	73.15	68.24	11.61 $\downarrow$

4.1. Comparing with SOTA Methods

Results on Ref-AVS Dataset. As shown in Tab. 1, we evaluate our method on an audio-language-visual task. With the Hiera_base+ backbone, our approach outperforms GAVS [50] by 5.2% in Jaccard for seen scenarios, demonstrating an enhanced ability to integrate complex multimodalities. Further, upgrading to the Hiera.L backbone yields an average Jaccard improvement of 4.12% compared to Hiera_base+, as detailed in the 'Mix' rows. Additionally, our method (AuralFuser) can be directly integrated into SAM [21], improving over TSAM [41] by 4.17% on the Seen average, highlighting its strong generalisation.

Results on AVS Datasets. In Tab. 2, we evaluate our approach on the AVSBench dataset under the audio-visual setting. With the Hiera_base+ backbone, our method surpasses adapter-based counterparts, achieving a 4.34% Jaccard gain over SAMA-AVS [30] on the V1m subset, highlighting the effectiveness of our cross-modal fusion design. Moreover, our method outperforms the zero-shot baseline [18] by 22.8%, demonstrating that our feature-level prompts provide stronger guidance to SAM2 than external foundation models. Our method also improves the SAM [21] architecture; for example, it boosts performance on the V1m subset by 2.12% over SAMA-AVS [30] and 1.52% over GAVS [50], effectively mitigating the audio prompt dilution issue across different SAM families.

4.2. Ablation Studies

We summarize component-wise performance gains in Tab. 3. The first row shows the baseline using only the visual modality. Incorporating audio and language in Ref-

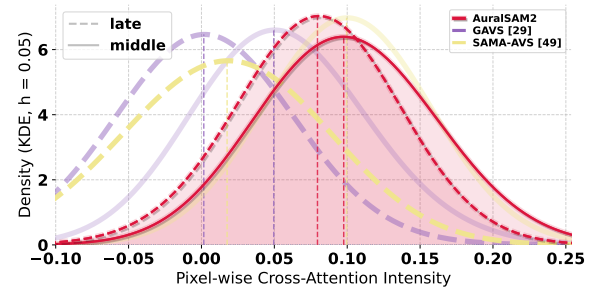


Figure 4. **PDF of cross-attention intensity** between audio cues and pixels on AVSBench (V1m) [58]. Values indicate pixel-wise cross-attention between the audio prompt and visual features.

AVS[51] improves $\mathcal{J} \& \mathcal{F}$ by 8.25% on AVSBench (V1m) and 9.54% on the Seen subset of Ref-AVS. Adding the feature pyramid further boosts performance by 3.55% and 2.57% on the respective datasets, demonstrating its effectiveness in capturing richer semantics for cross-modal fusion. Finally, introducing AudioCon improves results by another 1.25% and 0.84%, enhancing the alignment between vision and other modalities.

Probability Density of Cross-Attention Intensity across the network on AVSBench (V1m) [58], shown in Fig. 4. In later attention layers, our method exhibits a higher density centered around 0.075, while SAMA-AVS [30] peaks near 0.01, indicating that our approach effectively mitigates audio prompt dilution and enables stronger prompting. Notably, the shift in density modes between mid and late stages is smaller for our method, suggesting more consistent audio-visual alignment across the network propagation.

Ablation Studies on Feature-Level Prompts. As shown in Tab.5, we evaluate the importance of feature-level prompts by omitting them one at a time in Eq. (5) on AVSBench

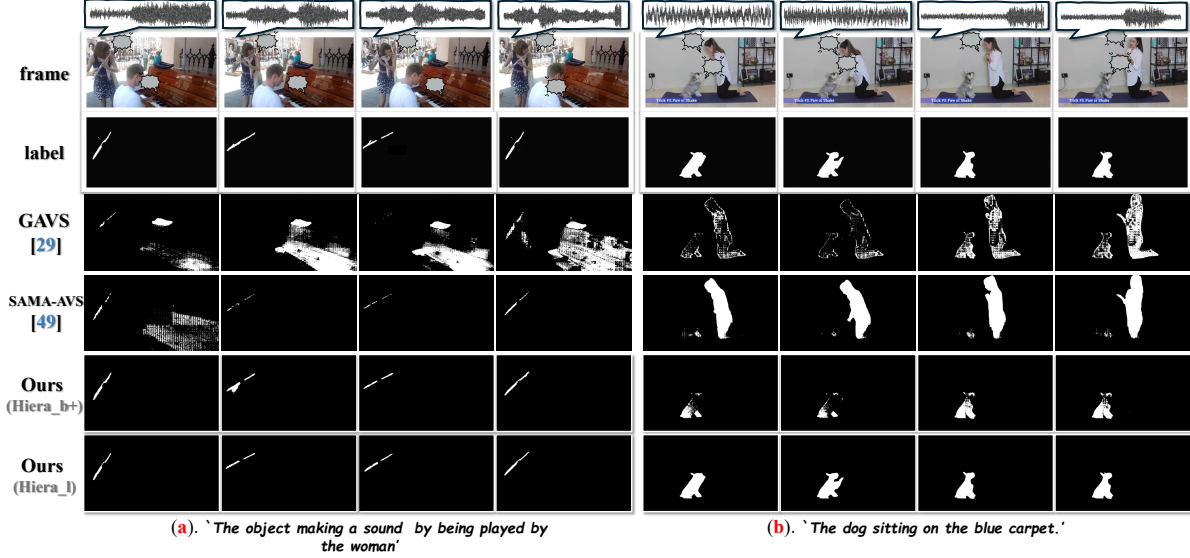


Figure 5. Qualitative visualisations on the Ref-AVS [51] dataset. The first row shows the input frame, followed by the ground truth labels in the second row. The third and fourth rows present adaptor-based methods [30, 50] using the SAM2 architecture with the Hiera.b+ backbone, while our method is displayed in the last two rows. Please refer to Supp. Section 4 for additional qualitative results.

Table 6. **Prompt Engineering with Audio** in the AVSBench (V1m) [58] dataset with Hiedra_base+ backbone. We use points and boxes generated from ground truth to simulate real-world prompting practices. The FPS represents the number of frames processed per second, and the best results highlighted in red.

Methods	Prompts	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	FPS
SAM2 [42]	points	64.67	72.15	17.8
	box	68.85	76.52	17.4
	mask	75.73	81.54	16.9
	points box	72.64	79.56	17.2
GAVS [50] (w/ SAM2)	audio points box	71.70	81.94	8.7
SAMA-AVS [30] (w/ SAM2)	audio points box	69.74	80.97	9.9
Ours (w/ SAM2)	audio points box	74.26	83.58	14.1

(V1m) with the Hiera.l backbone. The results indicate that both are essential to our module; for example, removing sparse prompts reduces the $\mathcal{J}\&\mathcal{F}$ score by 8.06%, while removing dense prompts decreases it by 11.61%.

Ablation Studies on CL. In Tab. 4, we present ablation studies on contrastive learning using the AVSBench (V1m) [58] dataset. The first row reports results without CL, the second row applies SupCon [22], originally designed for vision-only tasks, and the last row showcases our proposed AudioCon. Our method achieves an additional 0.77 $\mathcal{J}\&\mathcal{F}$ improvement over SupCon in AVS, demonstrating superior audio-visual alignment.

Promptable segmentation with SAM2. In Tab. 6, we simulate prompt engineering in a human-in-the-loop setting on AVSBench (V1m) [58]. The visual prompts are derived from the ground truth, consisting of four uniformly generated points per frame along with the corresponding bounding box, applied to the first frame following the SAM2 inference pipeline. Since preserving pixel-level labelled masks in practice is challenging, we

use only points and boxes in this experiment. As a result, compared to other adaptor-based methods [30, 50], our approach achieves the best performance in both measurements. For example, it increases Jaccard by 2.56% over GAVS [50] while maintains high efficiency with 14.1 frame-per-second (FPS) throughput.

4.3. Visualisation

We present qualitative results on Ref-AVS [51] in Fig. 5, where our method shows superior visual performance. In case (a), given the expression ‘the object making a sound by being played by the woman’, prior methods [30, 50] either misidentify the piano or fail to accurately segment the thick flute accurately. In contrast, our approach precisely captures the flute, achieving higher accuracy with the Hiera.l backbone.

5. Conclusion

We introduce AuralSAM2, a novel framework that enables SAM2 to process audio without relying on adapters or external foundation models. To address the inefficiency of repeated inference, we propose AuralFuser, a module that integrates multimodal features and directly generates sparse and dense feature-level prompts. These prompts guide the decoder without modifying image features, preserving SAM2’s efficiency and generalizability in promptable segmentation. To mitigate audio prompt dilution, AuralFuser performs cross-modal fusion within a multi-scale feature pyramid, enhancing both contextual understanding and fine-grained alignment. Finally, to alleviate visual dominance in the latent space caused by the imbalance between visual and audio embeddings, we introduce AudioCon, which promotes alignment around audio signals as semantic anchors.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3
- [2] Juan Leon Alcazar, Moritz Cordes, Chen Zhao, and Bernard Ghanem. End-to-end active speaker detection. In *European Conference on Computer Vision*, pages 126–143. Springer, 2022. 1
- [3] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018. 1
- [4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 2
- [5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 4
- [6] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 3
- [7] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Zero-shot audio source separation through query-based learning from weakly-labeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4441–4449, 2022. 3
- [8] Shengkai Chen, Yifang Yin, Jinming Cao, Shili Xiang, Zhen-guang Liu, and Roger Zimmermann. Openavs: Training-free open-vocabulary audio visual segmentation with foundational models. *arXiv preprint arXiv:2505.01448*, 2025. 1
- [9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26497–26507, 2024. 3, 4, 5
- [10] Yuanhong Chen, Kazuki Shimada, Christian Simon, Yukara Ikemiya, Takashi Shibuya, and Yuki Mitsufuji. Cstereo: Audio-visual contextual and contrastive learning for binaural audio generation. *arXiv preprint arXiv:2501.02786*, 2025. 2, 3
- [11] Yuanhong Chen, Chong Wang, Yuyuan Liu, Hu Wang, and Gustavo Carneiro. Cpm: Class-conditional prompting machine for audio-visual segmentation. In *European Conference on Computer Vision*, pages 438–456. Springer, 2025. 3, 4, 5
- [12] Shuangrui Ding, Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Yuwei Guo, Dahua Lin, and Jiaqi Wang. Sam2long: Enhancing sam 2 for long video segmentation with a training-free memory tree. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13614–13624, 2025. 1
- [13] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. *arXiv preprint arXiv:2307.01146*, 2023. 2, 6
- [14] Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. *arXiv preprint arXiv:2308.08288*, 2023. 3, 6
- [15] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Interactive video object segmentation using global and local transfer modules. In *European Conference on Computer Vision*, pages 297–313. Springer, 2020. 1
- [16] Yusuke Hosoya, Masanori Suganuma, and Takayuki Okatani. Rethinking annotation for object detection: Is annotating small-size instances worth its cost? *arXiv preprint arXiv:2412.05611*, 2024. 1
- [17] Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio visual segmentation. *arXiv preprint arXiv:2309.09501*, 2023. 3
- [18] Shaofei Huang, Rui Ling, Hongyu Li, Tianrui Hui, Zongheng Tang, Xiaoming Wei, Jizhong Han, and Si Liu. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. *arXiv preprint arXiv:2408.15876*, 2024. 1, 3, 6, 7
- [19] Shaofei Huang, Rui Ling, Tianrui Hui, Hongyu Li, Xu Zhou, Shifeng Zhang, Si Liu, Richang Hong, and Meng Wang. Revisiting audio-visual segmentation with vision-centric transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8352–8361, 2025. 3
- [20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2, 3
- [21] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 7
- [22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 8
- [23] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 3, 6
- [24] Yuma Koizumi, Yohei Kawaguchi, Keisuke Imoto, Toshiaki Nakamura, Yuki Nikaido, Ryo Tanabe, Harsh Purohit, Kaori Suefusa, Takashi Endo, Masahiro Yasuda, et al. Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring. *arXiv preprint arXiv:2006.05822*, 2020. 1

- [25] Jiaxu Li, Songsong Yu, Yifan Wang, Lijun Wang, and Huchuan Lu. Selm: Selective mechanism based audio-visual segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3926–3935, 2024. 3
- [26] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xun. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. *arXiv preprint arXiv:2309.09709*, 2023. 3
- [27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4, 5
- [28] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309, 2023. 3
- [29] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 1
- [30] Jinxiang Liu, Yu Wang, Chen Ju, Chaofan Ma, Ya Zhang, and Weidi Xie. Annotation-free audio-visual segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5604–5614, 2024. 1, 2, 3, 6, 7, 8
- [31] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 3
- [32] Tongkun Liu, Bing Li, Xiao Jin, Yupeng Shi, Qiuying Li, and Xiang Wei. Exploring few-shot defect segmentation in general industrial scenarios with metric learning and vision foundation models. *arXiv preprint arXiv:2502.01216*, 2025. 1
- [33] Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D Plumbley, and Wenwu Wang. Separate anything you describe. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024. 3
- [34] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364, 2019. 4
- [35] Juncheng Ma, Peiwen Sun, Yaoting Wang, and Di Hu. Stepping stones: A progressive training strategy for audio-visual semantic segmentation. *IEEE European Conference on Computer Vision (ECCV)*, 2024. 3, 6
- [36] Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Multimodal variational auto-encoder based audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 954–965, 2023. 3
- [37] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 218–234. Springer, 2022. 3
- [38] Shentong Mo and Yapeng Tian. Av-sam: Segment anything model meets audio-visual localization and segmentation. *arXiv preprint arXiv:2305.01836*, 2023. 3
- [39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5
- [40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2
- [41] Abduljalil Radman and Jorma Laaksonen. Tsam: Temporal sam augmented with multimodal prompts for referring audio-visual segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23947–23956, 2025. 6, 7
- [42] 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. 1, 3, 5, 6, 8
- [43] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 6
- [44] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pages 29441–29454. PMLR, 2023. 4, 7
- [45] Luisa Schwirten, Jannes Scholz, Daniel Kondermann, and Janis Keuper. Ambiguous annotations: When is a pedestrian not a pedestrian? *arXiv preprint arXiv:2405.08794*, 2024. 1
- [46] 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, 3
- [47] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. In *Asian Conference on Machine Learning*, pages 1215–1230. PMLR, 2024. 1
- [48] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2
- [49] Stéphane Vujasinović, Stefan Becker, Sebastian Bullinger, Norbert Scherer-Negenborn, Michael Arens, and Rainer Stiefelhausen. Strike the balance: On-the-fly uncertainty based user interactions for long-term video object segmentation. In *Proceedings of the Asian Conference on Computer Vision*, pages 2784–2802, 2024. 1
- [50] Yaoting Wang, Weisong Liu, Guangyao Li, Jian Ding, Di Hu, and Xi Li. Prompting segmentation with sound is generalizable audio-visual source localizer. In *Proceedings of*

- the AAAI Conference on Artificial Intelligence, pages 5669–5677, 2024. [1](#), [2](#), [3](#), [6](#), [7](#), [8](#)
- [51] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. In *European Conference on Computer Vision*, pages 196–213. Springer, 2025. [3](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [52] Yuji Wang, Haoran Xu, Yong Liu, Jiaze Li, and Yansong Tang. Sam2-love: Segment anything model 2 in language-aided audio-visual scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28932–28941, 2025. [6](#)
- [53] Xiaoli Wei, Zhaoqing Wang, Yandong Guo, Chunxia Zhang, Tongliang Liu, and Mingming Gong. Training-free robust interactive video object segmentation. *arXiv preprint arXiv:2406.05485*, 2024. [1](#)
- [54] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023. [1](#)
- [55] 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, 2023. [6](#)
- [56] Jiarui Yu, Haoran Li, Yanbin Hao, Jinmeng Wu, Tong Xu, Shuo Wang, and Xiangnan He. How can contrastive pre-training benefit audio-visual segmentation? a study from supervised and zero-shot perspectives. In *BMVC*, pages 367–374, 2023. [3](#), [6](#)
- [57] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. [5](#)
- [58] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 386–403. Springer, 2022. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#), [8](#)
- [59] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*, 2023. [3](#), [6](#), [7](#)
- [60] Jinxing Zhou, Yanghao Zhou, Mingfei Han, Tong Wang, Xiaojun Chang, Hisham Cholakkal, and Rao Muhammad Anwer. Think before you segment: An object-aware reasoning agent for referring audio-visual segmentation. *arXiv preprint arXiv:2508.04418*, 2025. [1](#), [2](#)
- [61] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18(3):351–376, 2021. [3](#)