Shift-Aware Identity-Guided Latent Refinement for Referring Audio–Visual Segmentation

Kun Li[1,3] **Sami Sebastian Brandt**[2] **Michael Ying Yang**[3,*]
[1] University of Liverpool, UK [2] IT University of Copenhagen, DK [3] University of Bath, UK
k.li51@liverpool.ac.uk, sambr@itu.dk, myy35@bath.ac.uk

Abstract

Referring Audio–Visual Segmentation (Ref-AVS) is an emergent multimodal task requiring the precise identification and segmentation of a specific object based on a natural language expression grounded in both visual and auditory cues. Unlike traditional AVS, which segments generic sound sources, or referring video object segmentation (Ref-VOS), which relies solely on visual–textual grounding, Ref-AVS necessitates fine-grained tri-modal interaction across vision, audio, and language. However, existing methods often fail to maintain consistency when subjected to target shift, a phenomenon where the model erroneously transfers its focus between objects due to transient occlusions, acoustic fading, or identity ambiguities. To address this, we propose a novel **S**hift-**A**ware **I**dentity-guided **L**atent (**SAIL**) refinement framework. It introduces a dual-token strategy that decouples frame-level dynamics from a global semantic anchor to explicitly model target-shift cues. We design a shift-aware refinement module that rectifies drift by aligning tokens with modality-specific evidence. Finally, we leverage a decoder that integrates the refined embeddings with a continuous memory update to maintain spatio-temporal coherence and identity consistency throughout the video sequence. Extensive experiments conducted on two Ref-AVS benchmarks demonstrate the effectiveness of our proposed method, significantly outperforming existing AVS, Ref-VOS, and Ref-AVS approaches. The code is released under `https://lik1996.github.io/SAIL/`.

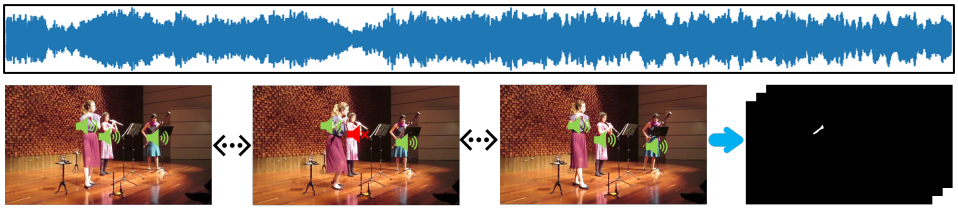
1 Introduction

The integration of visual and auditory modalities is fundamental to how humans perceive, interpret, and navigate the multifaceted complexity of dynamic real-world environments. Consequently, robust multimodal understanding has become a cornerstone for a broad spectrum of intelligent systems, ranging from event localization [40] and target tracking [14] to holistic scene understanding [25]. Building upon these foundations, Referring Audio–Visual Segmentation (Ref-AVS) [45] has emerged as a granular tri-modal task that requires not only understanding the language query provided by users but also precisely grounding a specific object within the spatio-temporal domain. While tasks such as audio–visual question answering (AVQA) focus on high-level linguistic reasoning [17, 22, 24], Ref-AVS demands a dense, pixel-level mask of a target identified through the synergistic alignment of visual, auditory, and textual cues.

Ref-AVS fundamentally differs from related multimodal segmentation tasks by requiring a dense, object-level grounding of *tri-modal* signals to resolve perceptual ambiguities, as shown in Fig. 1a. While Audio–Visual Segmentation (AVS) [8, 55] focuses on segmenting generic sound sources at the category level, Ref-AVS demands the resolution of fine-grained object-level identities, particularly in complex polyphonic scenes with multiple active sounds. Similarly, whereas Referring Video

*Corresponding author

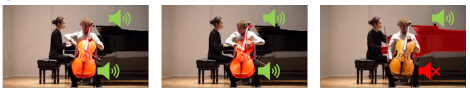
40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Referring Audio–Visual Segmentation (Ref-AVS) task.



(b) Target shift in the visual domain.



(c) Target shift in the auditory domain.

Figure 1: Illustration of Ref-AVS and target shift problems. (a) Given a tri-modal input (audio, video, and natural language expression), a Ref-AVS model performs sequential mask prediction for the referred object. (b) Visual target shift caused by transient occlusion or overshadowing (*e.g.*, shifting from the accordion to a guitar). (c) Auditory target shift caused by intermittent sound fading or pauses (*e.g.*, shifting from the cello to a piano).

Object Segmentation (Ref-VOS) [23, 47] relies exclusively on visual–textual grounding, Ref-AVS necessitates tri-modal alignment to handle scenarios where visual cues alone are insufficient for target identification. For example, in a scene featuring multiple violinists, only Ref-AVS can leverage a query like "the violin playing the staccato notes" to isolate the target by correlating specific auditory rhythms with localized visual motion. This intersection of modalities makes Ref-AVS a higher-level task, testing a model's ability to fuse disparate sensory streams into a singular, cohesive object representation, thereby achieving a finer-grained level of video understanding.

Despite the conceptual strength of tri-modal grounding, a critical yet under-explored challenge in Ref-AVS is the *target shift* problem. We define target shift as the temporal drift of predicted object identity throughout a video sequence. Formally, target shift occurs when the segmentation mask at time t corresponds to an object instance O' that differs from the target O referred to by the language query. Such shifts typically arise when visual or auditory cues become temporarily unreliable (*e.g.*, occlusion or sound absence), causing the model to re-anchor its prediction to another salient object. For instance, as illustrated in Fig. 1b, a target may be temporarily occluded or overshadowed by other entities; in such cases, the model's focus may erroneously shift from the referred accordion to a dominant visual distractor, such as another sounding instrument (*e.g.*, a guitar) that remains visually prominent. Similarly, a target's characteristic sound may intermittently fade or pause (*e.g.*, the cello and piano in Fig. 1c), triggering an auditory drift. This is especially common when an auditory distractor is explicitly mentioned in the textual guidance, misleading the model's focus. During these intervals of sensory ambiguity, maintaining a persistent identity is essential for a Ref-AVS model.

Existing Ref-AVS models fail to handle the target shift problem. Specialized multimodal approaches [34, 45] typically rely on localized spatial and temporal alignment to establish cross-modal correspondences. While effective in relatively static scenes, these methods struggle to maintain object identity across sensory gaps, leading to fragmented masks or inconsistent tracking. Recently, Multimodal Large Language Model (MLLM) based methods [6, 46] attempt to achieve tri-modal alignment through implicit reasoning. However, the generated global representation fails to capture fine-grained spatio-temporal details or preserve the target consistently during scene changes [10, 51]. This absence of a robust, persistent target identity mechanism leaves a significant gap in Ref-AVS. Furthermore, these MLLM approaches frequently depend on computationally expensive multi-stage training on additional datasets for instruction tuning [31].

To address the limitations, we propose **SAIL**, a novel end-to-end MLLM-based framework that leverages anchor identity and shift refinement for the Ref-AVS task. Specifically, we introduce a temporal `[seg]` token to capture frame-specific focus and a global `[query]` token to establish a persistent semantic anchor for the referred object. By explicitly modeling the variance across visual and auditory modalities, our framework quantifies potential target drift through a dedicated shift-estimation design. Finally, these shift-refined embeddings are integrated with a weighted

global anchor and then fed into a mask decoder. By synchronizing these refined features with a continuous memory bank update, SAIL effectively mitigates target shift and ensures spatio-temporal identity consistency for mask prediction. We validate the proposed method on two benchmarks Ref-AVSBench [45] and R^2-AVSBench [57], and the results show that SAIL significantly outperforms previous AVS, Ref-VOS, and Ref-AVS methods. Our **main contributions** can be summarized as:

- We identify and formalize the *target shift* problem in Ref-AVS, where transient cross-modal inconsistencies lead to unstable segmentation predictions across time.
- We propose SAIL, a novel end-to-end MLLM-based framework for Ref-AVS that explicitly models object identity and mitigates modality-induced drift via shift-aware refinement.
- We introduce an identity-guided token design with shift-aware update, which disentangles persistent anchor semantics from transient multimodal evidence and enforces consistent tri-modal grounding.
- Extensive experiments demonstrate that SAIL achieves new state-of-the-art performances on the Ref-AVSBench and R^2-AVSBench.

2 Related work

Referring Image Segmentation (RIS) serves as a foundational task for language-driven pixel-level understanding, where the goal is to generate the binary mask of a target described by a text. Existing RIS methods [5, 20, 21] primarily focus on effectively fusing visual and linguistic features. For example, LAVT [50] demonstrated the efficacy of early vision-language fusion within a transformer-based encoder. Referring Video Object Segmentation (Ref-VOS) extends this challenge into the temporal domain, requiring the model to localize the referred target while maintaining spatio-temporal consistency across multiple frames. URVOS [37] adopted a segment-then-track strategy, performing segmentation on key frames followed by mask propagation via cross-modal and memory attention mechanisms. ReferFormer [47] introduced language-conditioned object queries to track the referred objects across frames. With the emergence of foundation models [15, 33], recent approaches (*e.g.,* LISA [16] and VISA [49]) have leveraged MLLMs to achieve image- and video-level reasoning segmentation. However, these methods focus solely on visual–textual grounding, which is insufficient for handling the complex and diverse dynamics of audio–visual scenes.

Audio–Visual Segmentation (AVS) aims to segment the pixel-level regions of objects that are actively producing sound within a video sequence. Unlike traditional video segmentation, AVS requires the model to establish a fine-grained correspondence between audio signal and spatial pixels [26, 56]. The task was formally pioneered by AVSBench [55], which introduced a benchmark for both single-source and multi-source sounding object segmentation. AVSegFormer [8] utilized a transformer-based backbone to treat the audio signal as a global query, attending to visual features to decode the object masks. GAVS [44] and SAMA [27] have explored the integration of Segment Anything Model (SAM) [15] to leverage its powerful zero-shot spatial priors, allowing for more precise boundary delineation of sounding entities. By learning the visual representations of acoustic events, AVS provides a critical technical foundation for scene understanding, though it operates without the explicit high-level semantic guidance provided by natural language.

Referring Audio–Visual Segmentation (Ref-AVS) is a recently emerged task requiring pixel-level identification of targets specified by the synergistic alignment of auditory, visual, and linguistic cues. Wang *et al.* [45] established the foundational Ref-AVSBench dataset and proposed a baseline that adaptively combines tri-modal features for spatio-temporal correlation. Subsequent research has sought to enhance mask fidelity by incorporating foundation models. For instance, TSAM [34] integrated SAM to exploit robust spatial and temporal priors, while SAM2-LOVE [46] leveraged SAM2 [35] by mapping multimodal representations into a learnable `[seg]` token for implicit reasoning, further enhanced through token propagation and accumulation. These specialized multimodal methods prioritize the design of cross-modal fusion mechanisms to align the three modalities. Recently, several works [12, 31, 52] have shifted toward MLLMs to leverage their powerful reasoning capabilities. Crab [6] proposed a unified audio–visual scene understanding framework with explicit inter-task cooperation via interaction-aware LoRA [11] adapters. TGS-Agent [57] introduced an explainable think-ground-segment pipeline; it employed LLMs to generate textual descriptions that guide object-level grounding via Grounding-DINO [28], followed by mask generation through SAM2. However, existing methods often fail to account for the dynamic nature of audio–visual scenes where the referred object may undergo significant changes.

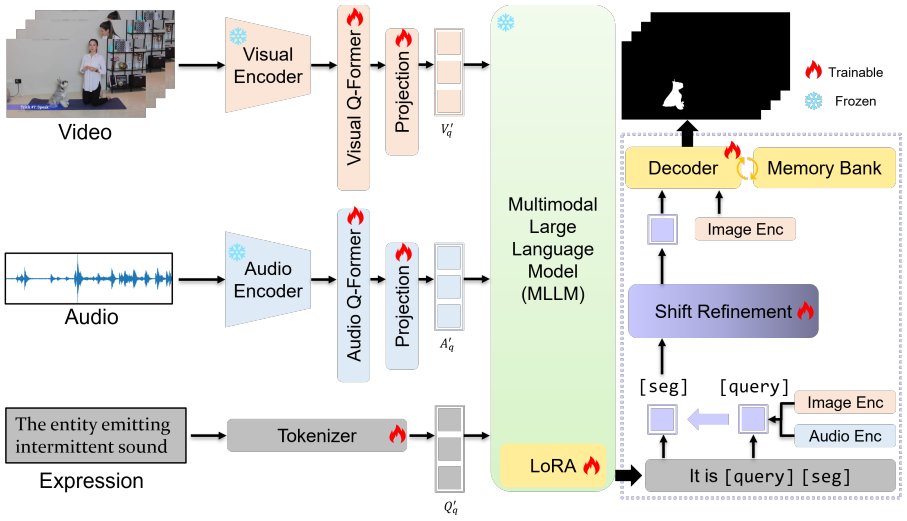


Figure 2: Overview of the proposed SAIL framework. Our MLLM with LoRA [11] reasons over the unified tri-modal tokens, generating frame-level `[seg]` tokens for local dynamics and a global `[query]` token as a semantic anchor. The shift-aware refinement module estimates and rectifies the latent shift by comparing the domain-specific evidence with the anchor's semantics. The mask decoder utilizes the refined embeddings and a memory mechanism to generate consistent mask predictions for the entire video sequence. For brevity, projection units before the decoder are omitted.

3 Method

3.1 Preliminary

Task formulation. The task of Ref-AVS can be defined as follows. Given a video sequence of T frames $\mathcal{V} \in \mathbb{R}^{T\times 3\times H\times W}$ and its synchronized audio track $\mathcal{A}$, along with a natural language query $\mathcal{Q}$, the objective is to learn a mapping function $\mathcal{F}_m : \{\mathcal{V}, \mathcal{A}, \mathcal{Q}\} \rightarrow M$ to generate a sequence of segmentation masks $M \in \mathbb{R}^{T\times H\times W}$ that identify the referent. Beyond grounding, the task requires the model to maintain identity permanence for the referred object across the entire video sequence.

Overall architecture. To address the challenges of Ref-AVS, we propose SAIL, a novel MLLM-based method with shift-aware refinement, thereby ensuring robust spatio-temporal identity consistency. Fig. 2 illustrates the overall architecture of the proposed method, which consists of three main components: i) an MLLM incorporating intra-modality feature tokenization in Sec. 3.2 and tri-modal alignment for dual-token generation in Sec. 3.3; ii) a shift-aware refinement module in Sec. 3.4; and iii) a memory-augmented mask decoder in Sec. 3.5. Note that the decoder is flexible and can be replaced with diverse models (more details can be found in *Appendix* A.4.4).

3.2 Input representation

Visual encoding. Given a T-second video sequence, we utilize a pre-trained CLIP-ViT-L/14 [33] encoder to extract patch-level features. For each frame, the encoder produces a feature map $f_v^t \in \mathbb{R}^{L_v\times D_v}$, where L_v and D_v represent the number of tokens and feature dimension, respectively. The entire video features can be denoted as $F_v = \{f_v^t\}_{t=1}^T$. To compress these high-dimensional feature maps, we employ Q-Former [18] with V_q learnable queries to represent them as a compact set of visual tokens. These tokens are subsequently projected into the LLM's input space via a two-layer multilayer perceptron (MLP).

Audio encoding. The corresponding audio stream is partitioned into T segments $\{a_t\}_{t=1}^T$ synchronized with the video frames. We leverage BEATs [1], a self-supervised audio transformer, to capture rich semantic information from each segment, resulting in audio embeddings $f_a^t \in \mathbb{R}^{L_a\times D_a}$, where L_a and D_a represent the number of tokens and feature dimension, respectively. Similar to the visual encoding, a dedicated Q-Former-based module compresses these features into A_q auditory queries. Following an MLP projection layer, these queries are aligned with the textual modality and prepended to the LLM's input sequence alongside the visual and linguistic embeddings.

3.3 MLLM for segmentation and dual-token generation

MLLM for Ref-AVS. Following the success of image-based reasoning segmentation models (*e.g.*, LISA [16]), recent Ref-AVS methods [6, 46, 52] leverage the implicit reasoning of LLMs to generate a specialized segmentation token, denoted as `[seg]`. In this paradigm, the MLLM processes the multimodal encodings to produce a high-level representation that drives a mask decoder Dec. A straightforward extension of this image-based approach can be formulated as:

$$[\texttt{seg}]_{t=1}^{N} = \mathrm{MLLM}(V_q', A_q', Q_q'), \quad M_{t=1}^{T} = \mathrm{Dec}(F_v, [\texttt{seg}]_{t=1}^{N}), \tag{1}$$

where V_q', A_q', and Q_q' represent the projected embeddings from the video, audio, and text, respectively, and N denotes the number of frames that an MLLM processes. The MLLM outputs a sequence of frame-specific tokens to represent the target at each time step. However, as previously analyzed, relying solely on independent `[seg]` tokens often fails to maintain a stable reference, making the model susceptible to target shift when sensory cues change.

Dual-token strategy. In contrast, we address the limitation by introducing a dual-token strategy designed to decouple dynamic per-frame focus from global semantic persistence. Our method prompts the LLM to generate two distinct types of tokens: (1) temporal `[seg]` tokens: a sequence of tokens $[\texttt{seg}]_t \in \{[\texttt{seg}]_1, \ldots, [\texttt{seg}]_N\}$, where each token is responsible for capturing the frame-specific tri-modal features for localized segmentation at time t; (2) global `[query]` token: a singular token that serves as a persistent semantic anchor. Unlike the `[seg]` tokens, the `[query]` token is optimized to summarize the long-term identity of the referred object across the entire video clip. By generating these tokens in a unified causal chain, the MLLM facilitates structured reasoning over tri-modal alignment across vision, audio, and language. It is under the standard paradigm using the same autoregressive cross-entropy loss as existing MLLM-based segmentation methods. Since the MLLM performs autoregressive decoding, placing `[query]` before `[seg]` naturally conditions the generation of the segmentation representation on the previously established target identity. The hidden states corresponding to the generated `[query]` and `[seg]` tokens are extracted separately. The `[seg]` tokens provide the flexibility to capture temporal dynamics, while the `[query]` token provides a constant reference point. This dual-token representation serves as the foundation for our subsequent shift-aware refinement, ensuring that even if a `[seg]` token drifts due to transient conflicts, the `[query]` anchor preserves the target's identity.

3.4 Shift-aware refinement module

To maintain robust, persistent segmentation under dynamic conditions, we propose the shift-aware refinement module, which bridges the gap between the LLM's latent tokens and the frame-level modality evidence, ensuring the mask decoder receives a dynamically calibrated prompt. Specifically, since raw visual and audio features reside in different modalities and often contain scene noise, we utilize the global semantic `[query]` to probe for target-specific evidence separately. By treating the `[query]` token as a persistent anchor, we compute cross-modal attention [41] to filter out irrelevant background information and isolate the referent:

$$z_m^t = \mathrm{Softmax}\left(\frac{(W_q[\texttt{query}])(W_k f_m^t)^T}{\sqrt{d}}\right) W_v f_m^t, \quad m \in v, a, \tag{2}$$

where z_m^t represents the target-specific evidence within each modality at frame t. Target shift is represented as the discrepancy between the LLM's current latent representation $[\texttt{seg}]_t$ and the actual observed evidence z_m^t. We compute the modality-specific shifts as:

$$\Delta_m = \mathrm{MLP}(\phi([\texttt{seg}]_t) - \psi(z_m^t)), \quad m \in v, a, \tag{3}$$

where ϕ and ψ are projection layers. These operations capture fine-grained visual motion or acoustic fluctuations that are inadequately represented by the LLM's discrete tokens, which tend to overlook subtle cues or overemphasize dominant modality signals. This estimation represents the "shift" because the deviation between the predicted semantic state and the synchronized sensory evidence signals a potential drift in focus, which the residual Δ_m explicitly quantifies. To prevent the model from being misled by dominant-modality changes or transient noise, we introduce dynamic confidence gating for the estimated drifts. Specifically, we compute the gating weights λ_v and λ_a based on the unnormalized attention logits generated during the evidence extraction stage. We derive the global evidence intensity, e_v and e_a, by averaging the pre-softmax alignment scores across the sequence

length for each modality. They preserve the absolute matching magnitude between the target and each modality, providing a more informative signal for controlling the refinement. Since these raw scores are not calibrated confidence values and may have different modality-dependent scales, we concatenate them and learn their joint calibration through a lightweight linear projection followed by a Sigmoid activation σ to obtain the gating weights, λ_v and λ_a. We apply these gated weights to ensure that the refinement prioritizes the modality providing the most reliable evidence at any given time step, effectively filtering out unreliable signals from modality-specific changes. It is formulated as:

$$\lambda_v, \lambda_a = \sigma(\text{Linear}([e_v; e_a])), \quad \hat{\texttt{[seg]}}_t = \texttt{[seg]}_t + \text{LN}(\lambda_v \Delta_v) + \text{LN}(\lambda_a \Delta_a), \tag{4}$$

where LN denotes LayerNorm. The resulting $\hat{\texttt{[seg]}}_t$ token incorporates both the stable semantic anchor and the dynamic sensory evidence, providing the decoder with shift-aware representations that are resilient to modality-specific drifts.

3.5 Mask decoding

Unlike previous methods that either treat frames independently through propagation [34] or employ standard SAM2 [35] as a generic temporal segmentation engine [46], our mask decoder incorporates the object anchor and estimated shifts to maintain consistent and persistent perception of the target over time. This step is a joint optimization of the memory bank from SAM2 and refined embeddings in an end-to-end manner. Specifically, to bridge the gap between the LLM and the mask decoder, we locate the tokens and project the corresponding last layer's hidden states into the embedding space of the SAM2 prompt encoder. Furthermore, we anchor the target with the `[query]` token alongside the temporal-level representations to enable inter-frame spatio-temporal consistency:

$$\mathcal{H}_t = \text{MLP}(\hat{\texttt{[seg]}}_t + \alpha\,\texttt{[query]}), \tag{5}$$

where $\mathcal{H}_t$ represents the prompt at frame t and α denotes the fusion coefficient. By utilizing the shift-refined tokens, this prompt inherently carries the corrected semantics, effectively guiding the decoder to focus on the intended referent. To ensure that the identity is preserved across the entire video sequence, we incorporate a continuous memory bank update. For each frame t, the decoder generates the mask M_t and simultaneously updates the temporal memory $\mathcal{M}$:

$$M_t = \text{Dec}(\text{Enc}_I(\mathcal{V}_t), \text{Enc}_P(\mathcal{H}_t), \mathcal{M}), \tag{6}$$

where Enc_I and Enc_P denote the SAM2 image encoder and prompt encoder, respectively. The shift-aware prompting and memory-based propagation further allow SAIL to achieve robust spatio-temporal identity consistency throughout the video timeline.

4 Experiments

4.1 Datasets and evaluation metric

Dataset. The experiments were primarily conducted on the widely-used Ref-AVSBench [45], a comprehensive dataset designed for Ref-AVS. It provides a dense mapping of 20,261 linguistic queries to pixel-level annotations across 4,002 diverse video sequences. The benchmark follows a standard split of 2,908 training, 276 validation, and 818 testing videos. Notably, the test set is further divided into three distinct scenarios: i) *Seen* (39 known categories); ii) *Unseen* (13 novel categories for testing out-of-distribution robustness); iii) *Null* (scenarios involving non-existent or invisible targets). We also evaluated models on R^2-AVSBench [57], an extension of Ref-AVSBench that incorporates reasoning-focused enhancements and comprises 400 testing videos (more details can be found in *Appendix* A.1).

Evaluation metric. Following the standard protocol in previous Ref-AVS studies [45, 46], we report the Jaccard index ($\mathcal{J}$) and F-score ($\mathcal{F}$) scores to quantify segmentation quality. For the *Null* subset, we report the S-score ($\mathcal{S}$), defined as the ratio of predicted mask area to total background area. This specialized metric assesses the model's ability to suppress false positives.

4.2 Implementation details

Videos were segmented into 1-second clips from 10-second recordings. Visual features were extracted using CLIP-ViT-L/14 [33], and audio features were captured via BEATs [1]. For both vision and

Table 1: Comparison with existing methods on Ref-AVSBench [45], where Mix represents the mean performance of Seen and Unseen subsets. † denotes methods requiring additional data for instruction tuning; ‡ indicates models that further necessitate external components in multiple stages. The best results are **bold** while the second best are underlined.

Method	Task	Model	Seen↑		Unseen↑		Mix↑		Null↓
			$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{S}$
AVSBench [55] *eccv22*	AVS	PVT-v2	23.2	51.1	32.4	54.7	27.8	52.9	0.208
AVSegFormer [8] *aaai24*		PVT-v2	33.5	47.0	36.1	50.1	34.8	48.6	0.171
GAVS [44] *aaai24*		SAM	28.9	49.8	29.8	49.7	29.4	49.8	0.190
SAMA [27] *wacv24*		SAM	39.2	56.2	47.5	56.6	43.4	56.4	0.130
ReferFormer [47] *cvpr22*	Ref-VOS	Video-Swin	31.3	50.1	30.4	48.8	30.9	49.5	0.176
R2-VOS [23] *iccv23*		Video-Swin	25.0	41.0	27.9	49.8	26.5	45.4	0.183
EEMC [45] *eccv24*	Ref-AVS	Mask2Former	34.2	51.3	49.5	64.8	41.9	58.1	**0.007**
TSAM [34] *cvpr25*		SAM	43.4	56.8	54.6	66.4	49.0	61.6	0.017
SAM2-LOVE [46] *cvpr25*		SAM2	43.5	51.9	66.5	72.3	55.0	62.1	0.230
Crab [6] † *cvpr25*		SAM + LLaMA2	40.5	-	45.6	-	43.1	-	-
OISA-1B [52] † *iccv25*		Mask2Former + InternVL2	51.7	58.7	58.3	65.1	54.5	61.4	0.098
TGS-Agent [57] ‡ *aaai26*		SAM2 + Groundin-DINO + LLaMA2	49.5	60.4	**73.2**	**80.6**	61.3	70.5	0.035
AURORA [31] ‡ *aaai26*		SAM + Qwen + VideoLLaMA2	63.2	72.8	69.7	76.4	66.5	74.6	-
SAIL (ours)	Ref-AVS	SAM2 + Chat-UniVi	**68.4**	**77.5**	72.0	79.5	**70.2**	**78.5**	0.010

audio modalities, we employed Q-Former [18] with 32 learnable query tokens to compress the features before projection. We adopted Chat-UniVi-7B [13] as our MLLM and performed fine-tuning using LoRA [11] with a rank of 8. We bypassed the need for auxiliary data for projector pre-training by jointly training the alignment layers and the memory-augmented SAM2 (Hiera-L) [35] decoder in a single, data-efficient end-to-end stage. We trained the model using AdamW [30] optimizer with an initial learning rate of $1e$-4 and a cosine decay schedule. All experiments were conducted using PyTorch on two NVIDIA H100 GPUs. More details are provided in *Appendix* A.1.

4.3 Comparison with state-of-the-art methods

Compared methods. To validate the effectiveness of the proposed method, we compared it with previous state-of-the-art approaches across three related tasks: AVS, Ref-VOS, and Ref-AVS. For AVS, we compared against AVSBench [55], AVSegFormer [8], GAVS [44], and SAMA [27]. We further adapted Ref-VOS methods that rely solely on text and vision, such as ReferFormer [47] and R2-VOS [23]. Finally, we compared against a comprehensive set of Ref-AVS methods, including specialized multimodal methods such as EEMC [45], TSAM [34], and SAM2-LOVE [46], as well as MLLM-based models including Crab [6], OISA-1B [52], TGS-Agent [57], and AURORA [31].

Quantitative results. The quantitative results on the Ref-AVSBench dataset are reported in Table 1. The specific components of these methods are listed in this table (further details are provided in *Appendix* A.1). Our proposed SAIL achieved state-of-the-art performances across most metrics, significantly outperforming existing AVS, Ref-VOS, and Ref-AVS methods. For instance, on the Mix split, SAIL surpassed the best-performing AVS baseline, SAMA, and representative Ref-VOS baseline, ReferFormer, by 26.8% and 39.3% in $\mathcal{J}$, respectively. These significant improvements underscore that simply extending single-modality or generic segmentation frameworks is insufficient for the complex tri-modal reasoning required in audio–visual scenarios. Notably, SAIL established a new performance ceiling for Ref-AVS. For instance, SAIL achieved 78.5% $\mathcal{F}$ on Mix, compared to 58.1% by the baseline EEMC and 61.4% by the omni-model OISA-1B. It also outperformed the previous SOTA method, AURORA, by 3.7% $\mathcal{J}$ and 3.9% $\mathcal{F}$. Detailed analysis shows that on the Seen split, our method achieved a 5.2% gain over AURORA in $\mathcal{J}$. These gains arise because SAIL explicitly addresses the target shift problem through token refinement rather than relying solely on the implicit reasoning of LLMs. Despite not leveraging additional instruction-tuning data (as used in OISA-1B) or complex multi-stage pipelines (as used in TGS-Agent and AURORA with external components such as Grounding-DINO [28] and Qwen [48]), SAIL still performed competitively on the Unseen split. It trailed the best-performing model, TGS-Agent, by only 1.1% in $\mathcal{F}$, while significantly outperforming other specialized multimodal methods (*e.g.,* SAM2-LOVE +7.2%) and MLLM-based methods (*e.g.,* AURORA +3.1%). The effectiveness of SAIL was further highlighted by the results on the Null split. SAIL achieved a remarkably low $\mathcal{S}$ score of 0.010, trailing only EEMC (0.007) but significantly outperforming other MLLM-based competitors like OISA-1B (0.098). This suggests that our method with the dual-token strategy effectively suppresses false positives caused by sensory noise.

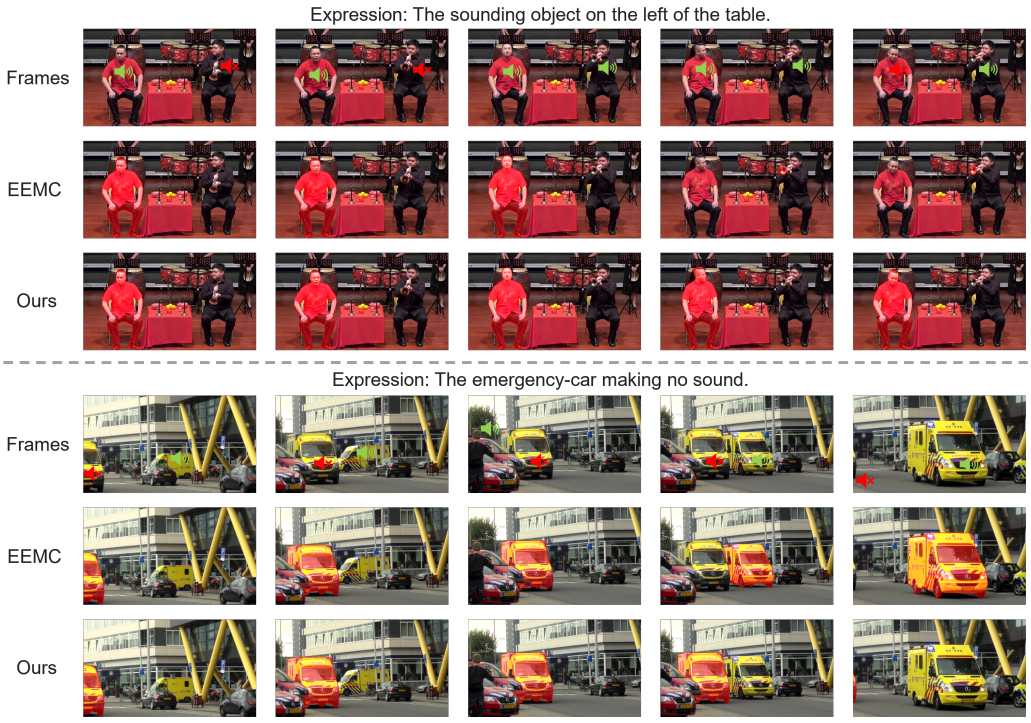


Figure 3: Qualitative results of SAIL and baseline from Ref-AVSBench [45].

To further validate the effectiveness of our method in addressing the target shift problem, we curated a specialized evaluation subset from the Ref-AVSBench test split, specifically targeting samples prone to significant temporal changes. Due to space constraints, a detailed description of this subset and an extended comparative discussion are provided in *Appendix* A.3.

Evaluation on R^2-AVSBench. To further assess the reasoning capabilities and generalizability of our model, we evaluated it on the R^2-AVSBench dataset. This benchmark replaces explicit object descriptions with implicit reasoning-heavy instructions, significantly increasing the difficulty of the task. The quantitative results are reported in Table 2. Compared to the results on Ref-AVSBench, we observe a general performance decline across all models, which is consistent with the increased complexity of interpreting implicit linguistic cues. Notably, SAIL maintained the inherent reasoning advantages of the MLLM-based architecture and achieved robust segmentation performance through our specialized anchor and shift-aware designs. Overall, SAIL achieved 64.4% in $\mathcal{J}\&\mathcal{F}$ on the Mix split, outperforming the previous SOTA method, TGS-Agent, by 4.3%. While TGS-Agent achieved higher performance on the Unseen split, it relied on a multi-stage pipeline with external components (*e.g.*, LLM-based reasoning and Grounding-DINO for explicit object localization), which may offer advantages in object grounding. In contrast, SAIL adopted a unified end-to-end framework without such auxiliary modules, while maintaining competitive performance and stronger robustness across the benchmark. These results indicate that even when linguistic inputs are implicit, our framework effectively grounds the referred object by leveraging the refined tri-modal evidence, demonstrating strong reasoning ability and generalization.

Qualitative results. In Fig. 3, we present qualitative results from the Ref-AVS test set. Overall, SAIL demonstrates superior target identification and segmentation precision. In the top example, while EEMC initially segmented the man performing vocal mimicry, it lost focus as the sound fades and other instruments (*e.g.*, suona) begin to play. In contrast, SAIL consistently identified the referent by leveraging shift-aware refinement to mitigate the impact of transient auditory changes. In the bottom example involving transient occlusion, EEMC erroneously tracked a distractor of similar appearance. SAIL successfully captured the latent shift even when the distractor hides behind and reappears near the true referent, effectively distinguishing between two similar emergency-cars. Notably, even when only a small portion of the target is visible in the final frame, our method maintained a correct and robust segmentation. These examples highlight SAIL's ability to leverage tri-modal dynamics and global anchoring, enabling accurate reasoning and segmentation in complex audio–visual scenes. Additional qualitative results and extended discussions are provided in *Appendix* A.5.

Table 2: Comparison on R^2-AVSBench [57]. Results for existing methods are sourced from either TGS-Agent or our re-implementations (in $*$). Note that TGS-Agent employed external Grounding-DINO [28] for explicit object localization. $\mathcal{J}\&\mathcal{F}$ represents the average of $\mathcal{J}$ and $\mathcal{F}$.

Method	Seen↑			Unseen↑			Mix↑		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
EEMC [45] *eccv24*	22.6	39.9	31.2	38.7	56.9	47.8	30.6	48.4	39.5
Grounded-SAM2 [36] *arxiv24*	21.4	30.1	25.7	44.9	53.4	49.1	33.2	41.7	37.4
Crab [6] *cvpr25*	19.6	32.1	25.9	36.9	52.0	44.4	28.2	42.1	35.2
TGS-Agent [57] *aaai26*	42.7	52.3	47.5	**68.3**	**77.0**	**72.7**	55.5	64.7	60.1
AURORA [31] $*$ *aaai26*	50.5	61.1	55.8	55.0	64.4	59.7	52.8	62.8	57.8
SAIL (ours)	**58.3**	**68.9**	**63.6**	60.2	70.1	65.2	**59.3**	**69.5**	**64.4**

Table 3: Ablation study on the main designs. The dual-token strategy, shift-aware refinement module, anchor alignment term and memory bank update are denoted as DK, SR, AA, and $\mathcal{M}$ respectively.

Method	Seen↑		Unseen↑		Mix↑		Null↓
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{S}$
Baseline	41.3	55.6	45.4	61.3	43.4	58.5	0.049
+DK	50.7	59.2	55.1	67.0	52.9	63.1	0.037
+DK+SR	61.8	69.7	68.3	75.5	65.1	72.6	0.019
+DK+SR+AA	64.6	73.8	70.2	77.1	67.4	75.5	0.015
+DK+SR+$\mathcal{M}$	65.0	75.2	70.4	78.0	67.7	76.6	0.013
SAIL (ours)	**68.4**	**77.5**	**72.0**	**79.5**	**70.2**	**78.5**	**0.010**

Table 4: Ablation study on the alternative design strategies (Pre-LLM, In-LLM, and Post-LLM) for tackling the target shift problem.

Method	Seen↑		Unseen↑		Mix↑		Null↓
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{S}$
Pre-LLM	43.8	51.7	58.0	69.5	50.9	60.6	0.075
In-LLM	66.3	74.1	59.7	66.2	63.0	70.2	0.042
Post-LLM	55.7	61.5	63.8	70.6	59.8	66.1	0.061
SAIL (ours)	**68.4**	**77.5**	**72.0**	**79.5**	**70.2**	**78.5**	**0.010**

4.4 Ablation studies

Ablation study on main components. To verify the effectiveness of the main components in SAIL, we conducted a cumulative ablation study on the Ref-AVSBench dataset. Table 3 reports the corresponding results, starting from a standard LISA-style MLLM baseline and progressively integrating our core designs. Removing all proposed designs resulted in a substantial performance drop of 20% in $\mathcal{F}$ on the Mix split, highlighting the overall importance of our framework. Replacing the single segmentation token with our dual-token strategy (*i.e.,* `[seg]` and `[query]`) yielded a significant boost, suggesting that decoupling frame-specific dynamics from a persistent semantic anchor helps the model maintain a consistent target identity. The introduction of the shift-aware refinement module provided the most substantial gain, increasing $\mathcal{J}$ and $\mathcal{F}$ by 12.2% and 9.5% on Mix, respectively. These sharp improvements validate our hypothesis that latent space drift, caused by modality noise or scene changes, is a primary bottleneck in Ref-AVS. By explicitly estimating and rectifying the shift between the LLM's tokens and target-specific evidence, the model achieved much higher grounding precision. Furthermore, aligning the weighted anchor with the segmentation representation (the anchor alignment term) and incorporating the SAM2 memory bank update (*i.e.,* $\mathcal{M}$) both independently enhanced temporal stability. For example, the anchor alignment improved $\mathcal{J}$ by 2.3% on the Mix split, while the memory bank update provided a 2.6% boost. These components ensure the referent is not "forgotten" during mask prediction. Finally, the complete SAIL framework, which integrates all modules, achieved peak performance across all splits and a minimum $\mathcal{S}$ score of 0.010 on the Null split, effectively suppressing false positives while maximizing segmentation accuracy. More detailed ablation studies on the modules and their alternatives can be found in *Appendix* A.4.

Ablation study on shift-aware refinement module. To verify the effectiveness of the proposed residual-based refinement and the dynamic gating mechanism, we conducted a structural ablation study. We compared our design against a simple fusion variance, where the segmentation token is updated via concatenation with modality evidence after projection, rather than calculating the identity drift. As shown in Table 5, replacing our residual-based drift estimation (*i.e.*, Δ) with a simple fusion strategy led to a significant performance drop (*e.g.*, $\mathcal{J}\&\mathcal{F}$ from 72.5% to 67.4% on the Mix split). This confirms the intuition that in the presence of target shift, simply "adding" modality features is insufficient. The residual design acts as a targeted error-correction signal, specifically identifying and rectifying the displacement between the current perceived evidence and the global identity anchor. We further concatenated the visual and auditory features before evidence computation, which consistently performed worse than our proposed SAIL. The inclusion of modality-specific gating weights (*i.e.*, λ) further enhanced the model's robustness, improving $\mathcal{J}\&\mathcal{F}$ by 1.9% on the Mix split.

Table 5: Ablation study on the shift-aware refinement module.

Fusion	Residual	Modality concate	Gating weights	Seen↑			Unseen↑			Mix↑			Null↓
				$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{S}$
✓				60.9	70.5	65.7	65.4	72.7	69.1	63.2	71.6	67.4	0.027
	✓			66.3	75.4	70.9	70.3	77.8	74.1	68.3	76.6	72.5	0.015
	✓	✓		63.5	72.6	68.1	69.4	76.0	72.7	66.5	74.3	70.4	0.015
	✓		✓	**68.4**	**77.5**	**73.0**	**72.0**	**79.5**	**75.8**	**70.2**	**78.5**	**74.4**	**0.010**

This demonstrates that the dynamic weights effectively suppress noisy residuals from unreliable modalities, such as audio segments with high background interference or visual frames with heavy occlusion, ensuring that only high-confidence evidence is used to refine the latent states. Overall, the combination of residual-based drift estimation and gating provided the most stable performance across both Seen and Unseen splits. This architecture ensures that SAIL not only captures the semantic identity but also accurately maintains the spatial localization even when the tri-modal signals are temporarily inconsistent. More detailed ablation studies on the proposed gating can be found in *Appendix* A.4.1.

Ablation on alternative design strategies for target shift. To further validate our shift-aware refinement designs, we compared it against three alternative strategies for tackling the target shift problem. Table 4 reports the corresponding results. (1) *Pre-LLM*: This strategy focuses on enriching the input space by providing the LLM with "object-centric" knowledge before it begins processing. Following standard practices in video object tracking and segmentation [7, 42], we extracted explicit object-level features via a pre-trained detector (*e.g.*, Mask R-CNN [9]). The $Top-k$ regions of interest (RoIs) were cropped and passed through a CLIP-ViT-L/14 encoder to obtain feature vectors. These vectors were projected to the LLM's hidden dimension via a two-layer MLP and prepended as a "visual" token to the prompt. While providing concrete visual priors, this strategy yielded the lowest performance (*e.g.*, 50.9% in $\mathcal{J}$), demonstrating that pre-LLM injection is insufficient for dynamic tri-modal reasoning. (2) *In-LLM*: Inspired by prompt-learning techniques [16, 49], we generated an implicit shift token alongside the segmentation token within the LLM. This strategy attempts to leverage the LLM's internal self-attention to manage shifts. Specifically, we expanded the LLM with a specialized `[shift]` token, while the hidden state was passed through a linear layer to predict an estimated shift. We then attempted a simple linear interpolation: $\mathcal{H}_t = \text{MLP}(\texttt{[seg]}_t + s\texttt{[shift]}_t)$, where s is a hyper-parameter. This strategy improved performance on the Seen split (74.1% in $\mathcal{F}$) but suffered on the Unseen split (66.2% in $\mathcal{F}$), suggesting that the LLM struggles to implicitly model spatial-temporal drift without explicit geometric grounding. (3) *Post-LLM*: This strategy explores whether the refinement should happen closer to the mask decoder than the reasoning LLM. Drawing from instructional editing models [53], we used a learnable query token to implicitly extract object-level dynamics from video features, which was then fused with the `[seg]` tokens. Without the global identity anchor guiding the attention, the model becomes overly sensitive to visual distractors. This strategy also failed to match SAIL, as this visual-heavy refinement often locks onto the wrong sounding object in Ref-AVS scenarios. Differently, by performing refinement in the latent space using an identity-guided anchor and shift-aware gating, our proposed SAIL maintains the semantic consistency of the target while learning multimodal changes over a video sequence. The results show that our SAIL framework achieved the best results across all splits. This confirms that rectifying latent shift with target-specific evidence is more effective than relying on explicit pre-processing or purely implicit token representations.

5 Conclusion

In this work, we explore the challenging problem of target shift in referring audio–visual segmentation and propose SAIL, a novel MLLM-based framework. To achieve robust and persistent segmentation in tri-modal video reasoning, we introduce a dual-token design that decouples frame-specific dynamics from a temporally consistent object identity. Building on this, we develop a shift-aware refinement module that explicitly estimates and rectifies modality-induced latent drift. By integrating the refined representations into a SAM2-based mask decoder, SAIL enables precise and temporally consistent segmentation. Extensive experiments on Ref-AVSBench and R^2-AVSBench demonstrate that SAIL consistently outperforms state-of-the-art methods, while ablation studies confirm the effectiveness of its components. We believe this work highlights the importance of identity-consistent representation learning and offers a promising direction for robust audio–video understanding.

Acknowledgments and Disclosure of Funding

This work has been supported by the Centre for Spatial Intelligence (RCSI) at the University of Bath, the European Union under grant agreement no. 101136006-XTREME, and the European Innovation Council under grant agreement no. 101257536-CEREBRIS. This work also benefited from computational resources provided by the Pioneer Centre for AI (DNRF grant number P1), Denmark, and the Barkla High Performance Computing facilities at the University of Liverpool, UK.

References

[1] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning*, pages 5178–5193, 2023.

[2] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[4] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[5] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7900–7916, 2022.

[6] Henghui Du, Guangyao Li, Chang Zhou, Chunjie Zhang, Alan Zhao, and Di Hu. Crab: A unified audio-visual scene understanding model with explicit cooperation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18804–18814, 2025.

[7] Yuqian Fu, Runze Wang, Bin Ren, Guolei Sun, Biao Gong, Yanwei Fu, Danda Pani Paudel, Xuanjing Huang, and Luc Van Gool. Objectrelator: Enabling cross-view object relation understanding across ego-centric and exo-centric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6530–6540, 2025.

[8] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12155–12163, 2024.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2969, 2017.

[10] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14623–14632, 2023.

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2):3, 2022.

[12] Dian Jin, Yanghao Zhou, Jinxing Zhou, Jiaqi Ma, Ruohao Guo, and Dan Guo. Simtoken: A simple baseline for referring audio-visual segmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 22702–22706, 2026.

[13] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024.

[14] Ben Kang, Xin Chen, Simiao Lai, Yang Liu, Yi Liu, and Dong Wang. Exploring enhanced contextual information for video-level object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4194–4202, 2025.

[15] 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.

[16] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[17] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022.

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742, 2023.

[19] Kun Li. Exemplar-guided interactive segmentation for multiple intra-class objects. *Neurocomputing*, page 133971, 2026.

[20] Kun Li, George Vosselman, and Michael Ying Yang. Scale-wise bidirectional alignment network for referring remote sensing image segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 226:350–363, 2025.

[21] Kun Li, Fu Wang, Feixiang Zhou, He Zhao, Xiaowei Xu, Yanda Meng, Yitian Zhao, and Yalin Zheng. Learning role-conditioned alignment for medical image referring segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 285–295, 2026.

[22] Kun Li, Michael Ying Yang, and Sami Sebastian Brandt. Query-guided spatial-temporal-frequency interaction for music audio-visual question answering. In *The Fourteenth International Conference on Learning Representations*, pages 81734–81752, 2026.

[23] Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, and Yan Lu. Robust referring video object segmentation with cyclic structural consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22236–22245, 2023.

[24] Zhangbin Li, Jinxing Zhou, Jing Zhang, Shengeng Tang, Kun Li, and Dan Guo. Patch-level sounding object tracking for audio-visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5075–5083, 2025.

[25] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1417–1424, 2013.

[26] Chen Liu, Peike Li, Hu Zhang, Lincheng Li, Zi Huang, Dadong Wang, and Xin Yu. Bavs: bootstrapping audio-visual segmentation by integrating foundation knowledge. *IEEE Transactions on Multimedia*, 26: 10015–10028, 2024.

[27] 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.

[28] 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, 2024.

[29] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[31] Ziyang Luo, Nian Liu, Fahad Shahbaz Khan, and Junwei Han. Aurora: Augmented understanding via structured reasoning and reinforcement learning for reference audio-visual segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 7716–7724, 2026.

[32] Trong-Thuan Nguyen, Pha Nguyen, and Khoa Luu. Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18384–18394, 2024.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[34] 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.

[35] 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. In *The Thirteenth International Conference on Learning Representations*, 2025.

[36] 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.

[37] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European Conference on Computer Vision*, pages 208–223, 2020.

[38] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[40] Sung Jin Um, Dongjin Kim, Sangmin Lee, and Jung Uk Kim. Object-aware sound source localization via audio-visual scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8342–8351, 2025.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[42] Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. Object-centric video question answering with visual grounding and referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22274–22284, 2025.

[43] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8 (3):415–424, 2022.

[44] 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*, volume 38, pages 5669–5677, 2024.

[45] 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, 2024.

[46] 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.

[47] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022.

[48] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[49] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pages 98–115, 2024.

[50] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[51] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. Ctvis: Consistent training for online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 899–908, 2023.

[52] Kaining Ying, Henghui Ding, Guangquan Jie, and Yu-Gang Jiang. Towards omnimodal expressions and reasoning in referring audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22575–22585, 2025.

[53] Shoubin Yu, Difan Liu, Ziqiao Ma, Yicong Hong, Yang Zhou, Hao Tan, Joyce Chai, and Mohit Bansal. Veggie: Instructional editing and reasoning video concepts with grounded generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15147–15158, 2025.

[54] Yian Zhao, Kehan Li, Pengchong Qiao, Chang Liu, Rongrong Ji, and Jie Chen. Unified granularity controller for interactive segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4051–4062, 2026.

[55] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403, 2022.

[56] 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. *International Journal of Computer Vision*, 133(4):1644–1664, 2025.

[57] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 13665–13673, 2026.

A Appendix

This supplementary document provides comprehensive details regarding the datasets and model architectures (Sec. A.1). We further present an analysis on target shift (Sec. A.2) and an expanded evaluation of target shifts (Sec. A.3), additional ablation studies (Sec. A.4), and more qualitative visualizations (Sec. A.5). In addition, we also provide complexity analysis (Sec. A.6) and discussion (Sec. A.7) for potential limitations and future work.

A.1 More details on datasets and models

A.1.1 Reasoning capability and dataset

While Ref-AVSBench [45] establishes the foundational benchmark for Ref-AVS, its linguistic queries primarily rely on explicit descriptions (*e.g.*, "The person playing the instrument"). In contrast, R^2-AVSBench [57] is specifically designed to evaluate more sophisticated cognitive abilities by introducing implicit reasoning challenges. The core distinction of R^2-AVSBench lies in its query formulation. Leveraging Gemini [38], the dataset replaces direct nouns with functional or acoustic descriptions, significantly increasing lexical and structural diversity. For example, a simple query in Ref-AVSBench such as "The dog barking in the yard" is transformed in R^2-AVSBench into "The creature making a repetitive high-pitched alert sound while standing on the grass." This change requires the model to not only localize the object but to first reason about the relationship between the acoustic signatures and the visual entities.

Consequently, R^2-AVSBench provides an extended evaluation set (400 in total: 220 and 180 videos for Seen and Unseen splits) derived from the Ref-AVSBench test split. The increased complexity is reflected in the linguistic statistics: the average query length in R^2-AVSBench is 11.73 words, a significant expansion compared to the 7.08 words found in the original benchmark. The experimental results reported in the main text (Sec. 4.3) further validate the superior reasoning capabilities of our proposed framework in handling complex, implicit Ref-AVS scenarios.

A.1.2 Architectural details of compared methods

In this section, we provide extended technical details for the baseline and state-of-the-art methods discussed in the main text (Sec. 4.3), focusing on their underlying architectural paradigms:

- AVSBench [55]. It employs a transformer-based architecture with a PVT-v2 [43] backbone and introduces a temporal pixel-level grouping strategy to associate audio features with spatial visual maps for sounding object localization.
- AVSegFormer [8]. Built on the PVT-v2 backbone, this model utilizes a transformer decoder where audio cues are injected into visual queries via cross-attention to amplify global audio–visual context and refine mask boundaries.
- GAVS [44]. It integrates the Segment Anything Model (SAM [15]) into the AVS task using a semantic-aware audio prompting strategy, guiding the foundation model to focus specifically on sounding regions.
- SAMA [27]. It is an evolution of SAM-based AVS that introduces multimodal attention. It prompts SAM with localized audio–visual fused features to leverage its high-fidelity mask generation capabilities.
- ReferFormer [47]. It is a unified framework for Ref-VOS using a Video-Swin [29] backbone. It treats linguistic expressions as object queries that directly attend to the most relevant spatial regions across video frames.
- R2-VOS [23]. Utilizing Video-Swin, this method introduces a structural text-to-text cycle consistency constraint. It jointly models segmentation and text reconstruction to enforce semantic consensus between video–text pairs.
- EEMC [45]. Built on Mask2Former [3], this model proposes an expression enhancement module that mines multimodal cues. It focuses on integrating dynamic audio–visual information directly into the reference expressions to resolve alignment.
- TSAM [34]. It is an end-to-end SAM-based architecture that redefines visual prompts as data-driven multimodal prompts. It incorporates a temporal modeling branch with sparse and dense prompting modules to stabilize segmentation.
- SAM2-LOVE [46]. It is the first adaptation of SAM2 [35] for Ref-AVS. It utilizes a multimodal fusion module to compress all the modalities' information into a learnable token, aligning with the SAM2 for segmenting the referred target.
- Crab [6]. It combines LLaMA-2-7b-chat [39] with SAM and introduces an interaction-aware LoRA [11] structure. It leverages LLM reasoning to interpret complex instructions but requires extensive auxiliary instruction-tuning data.
- OISA-1B [52]. It employs InternVL2-1B [2] and Mask2Former, emphasizing large-scale vision–language pre-training to improve open-vocabulary recognition and handle omnimodal reasoning tasks.
- TGS-Agent [57]. It is a modular "Think-Ground-Segment" pipeline coordinating LLaMA-2-7b-chat, Grounding-DINO [28], and SAM2. It uses the LLM as an agent to manage independent expert models for explicit description generation, grounding, and segmentation.
- AURORA [31]. It is a heavyweight ensemble integrating Qwen-Omni [48], VideoLLaMA2 [4], and SAM. While achieving high performance through a three-stage training procedure and massive scale, it relies on complex external components and auxiliary datasets.

A.1.3 More implementation details

Due to space constraints in the main text, additional implementation details are provided here. High-resolution frames were passed to SAM2 and augmented as described in [35], resized to 1024×1024. To explicitly model the target shift and maintain temporal identity, we constrained the MLLM to respond with a structured dual-token sequence: *It is* `[query] [seg]`. By placing `[query]` before `[seg]` in the causal chain of the LLM, the model ensures that the segmentation process is conditioned on a pre-established semantic identity. This design allows SAIL to effectively calculate and rectify the target shift within the latent space prior to mask decoding. For each time step t, we encoded the refined prompt and current video frame to SAM2. The resulting spatio-temporal features were processed by a memory encoder and stored in a memory bank as persistent anchors for future frames. We adopted the memory attention from SAM2, where the model performs cross-attention between the decoded features and the historical embeddings retrieved from the memory bank. The final input

to the mask decoder was a fusion of the current image features, the memory-attended context, and the refined embeddings. This ensures that the generated masks are not only spatially precise but also temporally consistent with the target.

We trained SAIL in an end-to-end manner using a multi-task loss function that optimizes both the linguistic reasoning of the MLLM and the pixel-level prediction of the mask decoder. Following the standard training protocol for MLLM-based segmentation (*e.g.*, LISA [16]), we employed an auto-regressive cross-entropy loss, $\mathcal{L}_{txt}$, for text generation. For the binary mask prediction, we calculated the discrepancy between the predicted masks M_t and the ground truth masks $\hat{M}_t$ across all T frames. The segmentation loss $\mathcal{L}_{seg}$ consists of a binary cross-entropy (BCE) loss and a DICE loss. The final objective function $\mathcal{L}$ is defined as:

$$\mathcal{L} = \lambda_{txt}\mathcal{L}_{txt} + \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice}, \tag{7}$$

where λ_{txt}, λ_{bce}, and λ_{dice} are hyper-parameters. The model was trained for 10 epochs with a batch size of 2 per device. The hyper-parameters λ_{txt}, λ_{bce}, λ_{dice} and α were set to 1.0, 1.0, 0.5 and 0.1, respectively.

A.2 Concrete causes of target shift

We believe the root cause of target shift is not insufficient multimodal fusion itself, but the absence of an explicit mechanism to preserve and verify target identity throughout multimodal reasoning. We argue that Ref-AVS consists of two complementary objectives: 1) identifying the referred target from multimodal evidence, and 2) maintaining this target identity while producing frame-wise segmentation under dynamic scenes. Existing specialized multimodal methods primarily establish local visual–audio correspondences on a frame-by-frame basis. As a result, their attention is dominated by the most salient multimodal evidence in each frame and may shift under inconsistent or ambiguous cues, leading to identity drift. Recent MLLM-based methods perform global tri-modal reasoning, yet their final `[seg]` representation is generated through implicit multimodal aggregation without explicitly assessing whether it remains consistent with the intended target. The resulting latent representation encodes relevant information for mask decoding, yet no explicit identity verification is performed during propagation. Consequently, once the `[seg]` representation drifts toward a distractor, the error propagates to subsequent mask prediction without any mechanism for detection or correction.

We provide an analysis to directly characterize target drift. Specifically, we measured the temporal consistency of the `[seg]` representation by Crab [6]. Since this representation encodes the model's understanding of the referred object, its temporal consistency reflects whether the target identity is preserved over time. For each video, we uniformly sampled a sequence of frames (*e.g.,* frames 1, 3, 5, 7, and 9) and computed the drift score between neighboring sampled frames as: $1 - cos(h_t, h_{i+1})$. We intentionally compared adjacent sampled frames, rather than the first and last frames, because videos inherently contain appearance and motion changes that may lead to gradual representation variations even when the correct target is maintained. Consequently, normal cases are expected to exhibit stable drift scores, whereas target shift cases should present a sudden increase once the latent representation deviates toward a distractor. As shown in Table 6, the two normal cases maintain consistently low drift scores throughout the video, while the target shift cases exhibit pronounced spikes (*e.g.,* 1.32 between Frames 3–5 in Drift-C). These observations also provide direct evidence, supporting our problem formulation.

Sample	Frames 1-3	Frames 3-5	Frames 5-7	Frames 7-9
Normal-A	0.13	0.11	0.27	0.16
Normal-B	0.22	0.48	0.14	0.28
Drift-C	0.20	1.32	0.34	0.15
Drift-D	0.71	0.15	1.07	0.09

Table 6: Drift score between neighboring sampled frames achieved by Crab.

Instead of attempting to further strengthen multimodal alignment, SAIL first maintains a persistent semantic anchor through the `[query]` token, then extracts target-specific evidence from each modality, explicitly estimates the discrepancy between this evidence and the current `[seg]` representation, and finally rectifies the segmentation state through shift-aware refinement. Consequently, SAIL is not merely compensating for alignment errors; it continuously verifies whether the predicted segmentation state remains consistent with the referred target throughout the learning process. This identity verification and correction mechanism enables SAIL to

Sample	Frames 1-3	Frames 3-5	Frames 5-7	Frames 7-9
Drift-A	0.14(0.17)	0.22(0.18)	0.18(0.14)	1.35(0.37)
Drift-B	0.30(0.21)	0.26(0.20)	1.46(0.53)	0.18(0.26)

Table 7: Additional diagnostic analyses on SAIL.

Table 8: Comparison with existing methods on the selective samples for target shift from Ref-AVSBench [45].

Method	Seen↑			Unseen↑			Mix↑		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
EEMC [45]	17.4	30.5	24.0	34.2	52.6	43.4	25.8	41.6	33.7
Crab [6]	25.1	40.7	32.9	32.3	50.5	41.4	28.7	45.6	37.2
TGS-Agent [57]	37.5	47.9	42.7	56.7	64.2	60.5	47.1	56.1	51.6
AURORA [31]	40.7	50.0	45.4	52.6	61.9	57.3	46.7	56.0	51.4
SAIL (ours)	**53.1**	**62.4**	**57.8**	**65.7**	**72.8**	**69.3**	**59.4**	**67.6**	**63.5**

remain effective regardless of whether the discrepancy originates from imperfect modality alignment, accumulated errors during sequential reasoning, or the entanglement of identity and frame-wise representations.

To further substantiate our claim, we conducted additional diagnostic analyses to characterize the target shift phenomenon. Specifically, we compared the evolution of the `[seg]` representations before and after the proposed refinement across video frames, allowing us to directly examine how the segmentation state changes during reasoning. In addition, we performed the same temporal consistency analysis on the `[seg]` representations before and after refinement in SAIL. Values are reported before (and after) refinement inside parentheses. As shown in Table 7, the proposed refinement suppressed the abrupt representation drift (*e.g.,* 1.35→0.37 and 1.46→0.53), indicating that the refined `[seg]` representation remains substantially more consistent across neighboring frames. These observations provide further evidence that SAIL effectively preserves target identity by explicitly correcting representation drift, rather than merely improving multimodal alignment.

A.3 Target shift analysis via specialized evaluation subset

To validate the effectiveness of our method in addressing the target shift problem, we curated a specialized evaluation subset derived from the Ref-AVSBench [45] test split. This subset focuses on challenging scenarios where the referred object undergoes transient changes in the visual or auditory domains, testing the model's ability to maintain stable consistency. This evaluation subset (Ref-AVSBench-TF) will be made publicly available.

Evaluation set. The selection process consists of two primary stages: automated visual filtering and manual auditory verification. We observed that in Ref-AVSBench, the target object is consistently present and annotated in the initial frame ($t = 0$). We leveraged this as a reference point to detect potential occlusions or identity drifts. Instead of a direct IoU comparison against the first frame, which can be misleading due to camera motion or natural target movement, we calculated the relative shrinkage ratio (R_s) between adjacent frames. When R_s falls below a predefined threshold (*e.g.*, 0.3), it signified a sudden disappearance or significant reduction of the target mask, indicating a potential occlusion or a distractor-induced shift. While the first step identifies visual disruptions, target shifts are often triggered by auditory complexity or scene density. We manually inspected the test split to include samples that might have been missed by the geometric filter, specifically targeting scenarios with intermittent sound changes, background sounding objects, or multiple simultaneous sound sources (*e.g.*, two people talking or overlapping instruments). We also checked videos containing more than one salient object of the same category, as these possess a higher probability of identity confusion during temporal propagation. Following this two-step pipeline, we identified 371 video sequences that exhibit significant visual or auditory shifts. Around 45% (371/818) of the test samples from Ref-AVSBench exhibit target shift, indicating that it is a prevalent challenge. This target shift evaluation set provides a focused set to verify the effectiveness of our proposed method on target shift problem in Ref-AVS.

Corresponding results. The results in Table 8 reveal a significant performance drop for all baseline methods compared to their results on the full Ref-AVSBench in Table 1. For instance, EEMC struggled severely on the Seen split (*e.g.*, 17.4% in $\mathcal{J}$), likely because its perception-heavy approach cannot recover once a visual target is temporarily lost. MLLM-based methods like TGS-Agent and AURORA performed better due to their stronger reasoning backbones, yet they still exhibited substantial drift. TGS-Agent achieved 47.1% in $\mathcal{J}$ on the Mix split, but failed to maintain the high performance (61.3%) for the complete test set. In contrast, SAIL demonstrated a significant performance advantage, outperforming AURORA by 12.7% in $\mathcal{J}$ on the Mix split. Specifically, on the Unseen split, where the model cannot rely on category-specific priors, SAIL maintained a high $\mathcal{F}$

Table 9: Ablation study on gating weights. We replaced the proposed dynamic modality-specific gating with two alternatives.

Method	Seen↑			Unseen↑			Mix↑			Null↓
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{S}$
No Gating	66.3	75.4	70.9	70.3	77.8	74.1	68.3	76.6	72.5	0.015
Shared Gating (λ)	68.2	77.0	72.6	71.4	78.5	75.0	69.8	77.8	73.8	0.011
Static Gating (β & γ)	67.3	76.9	72.1	70.6	78.1	74.4	69.0	77.5	73.3	0.014
Ours (λ_v & λ_a)	**68.4**	**77.5**	**73.0**	**72.0**	**79.5**	**75.8**	**70.2**	**78.5**	**74.4**	**0.010**

Table 10: Ablation study on LMMs. We replaced the MLLM in our SAIL with LLaMA2 [39] and InternVL2 [2], respectively.

Method	Seen↑			Unseen↑			Mix↑			Null↓
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{S}$
LLaMA2-7B [39]	65.1	73.7	69.4	69.2	76.7	73.0	67.2	75.2	71.2	0.028
InternVL2-8B [2]	67.6	76.8	72.2	70.6	77.4	74.0	69.1	77.1	73.1	0.014
Chat-UniVi-7B [13]	**68.4**	**77.5**	**73.0**	**72.0**	**79.5**	**75.8**	**70.2**	**78.5**	**74.4**	**0.010**

of 72.8%. This proves that the `[query]` token effectively acts as a universal semantic anchor that identifies the target even when visual features become noisy or novel. The consistent lead of SAIL across both Seen and Unseen scenarios in this specialized subset provides empirical evidence that our dual-token strategy and shift-aware latent refinement are not merely marginal improvements. Instead, they represent a fundamental architectural solution to the target shift problem of identity consistency in tri-modal audio–visual scene understanding.

A.4 More ablation studies

A.4.1 Ablation on gating weights

The effectiveness of the gating mechanism has already been partially validated in Table 5, where removing the gating module led to a noticeable performance drop (-1.9% on Mix). We further conducted two ablation studies. Shared dynamic gating predicts a single input-dependent weight λ from $[e_v; e_a]$ and applies it to both corrections, evaluating whether the two modalities require independent regulation. Static learnable gating replaces the evidence-conditioned weights with two learned scalars, β, γ, retaining modality-specific weighting while removing frame-dependent adaptation. This isolates whether the improvement comes from dynamic evidence estimation or merely from introducing additional learnable scaling parameters. As shown in Table 9, both variants improved over removing the gating mechanism, confirming the importance of adaptive weighting. However, both remained inferior to the proposed dynamic modality-specific gating. These results demonstrate that the performance gain comes not only from introducing learnable gating parameters, but also from estimating confidence dynamically and modulating visual and audio corrections independently, validating the design of our confidence gating mechanism.

A.4.2 Ablation on LLMs

To verify that the performance gains of SAIL are rooted in our proposed shift-aware refinement rather than the specific choice of the LLM backbone, we ablated the MLLM component. We replaced the Chat-UniVi [13] with LLaMA2 [39] (a text-centric LLM) and InternVL2 [2] (a high-capacity vision-language model). The results in Table 10 indicate that SAIL maintained strong performance regardless of the underlying MLLMs. Even with LLaMA2, the model achieved a $\mathcal{J}$ of 67.2% on the Mix split, compared to Crab (43.1%) that also used the LLM. The model with Chat-UniVi yielded the best results across all metrics. Unlike InternVL2, which is optimized for static images, Chat-UniVi is specifically designed for unified visual representation across images and videos. Its ability to handle temporal sequences more fluidly provides a more stable initial tokens for our refinement module. LLaMA2's lower performance on the Null split highlights the disadvantage of using a model that lacks integrated multimodal pre-training. The higher "Null" split error indicates that without a strong multimodal foundation, the model is more prone to hallucinating segments when no sounding object is present. In summary, while a stronger MLLM backbone can provide a better starting point, the consistent success of SAIL across different architectures confirms that our dual-token latent refinement is the primary driver of temporal consistency and target shift rectification.

Table 11: Ablation study on the fusion coefficient α in the anchor alignment before the mask decoder.

Method	Seen↑			Unseen↑			Mix↑			Null↓
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{S}$
$\alpha = 0$	65.0	75.2	70.1	70.4	78.0	74.2	67.7	76.6	72.2	0.013
$\alpha = 0.1$	**68.4**	**77.5**	**73.0**	**72.0**	**79.5**	**75.8**	**70.2**	**78.5**	**74.4**	**0.010**
$\alpha = 0.2$	67.2	75.0	71.1	70.8	78.2	74.5	69.0	76.6	72.8	0.011
$\alpha = 0.5$	66.7	74.3	70.5	70.1	77.6	73.9	68.4	76.0	72.2	0.012
$\alpha = 1.0$	63.4	72.4	67.9	68.5	76.7	72.6	66.0	74.6	70.3	0.017

Table 12: Ablation study on the mask decoder. QP denotes the query propagation operation in OISA-1B [52].

Method	Seen↑			Unseen↑			Mix↑			Null↓
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{S}$
+ M2F	62.8	71.9	67.4	65.8	73.2	69.5	64.3	72.6	68.5	0.020
+ SAM	64.3	73.8	69.1	67.7	75.0	71.4	66.0	74.4	70.2	0.015
+ QP + SAM	66.9	76.1	71.5	71.0	78.5	74.8	69.0	77.3	73.2	0.010
+ SAM2	65.0	75.2	70.1	70.4	78.0	74.2	67.7	76.6	72.2	0.013
SAIL(ours)	**68.4**	**77.5**	**73.0**	**72.0**	**79.5**	**75.8**	**70.2**	**78.5**	**74.4**	**0.010**

A.4.3 Ablation on fusion coefficient

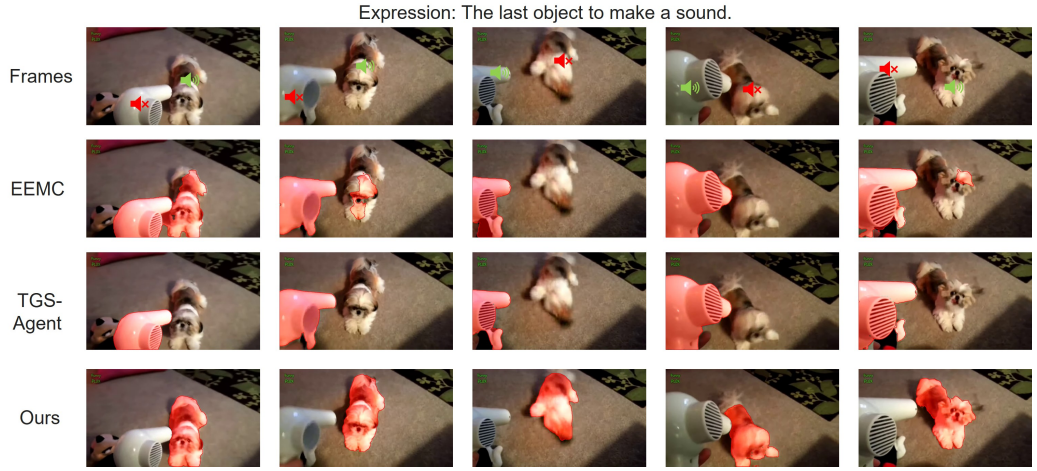
We investigated the sensitivity of the SAIL framework to the fusion coefficient α on Ref-AVSBench [45], which governs the intensity of the identity-guided refinement applied to the latent segmentation tokens. As shown in Table 11, the model achieved peak performance across all metrics when α is set to 0.1. The results reveal a consistent performance degradation as the value of α increases. Specifically, increasing α from 0.1 to 1.0 led to a notable drop in $\mathcal{J}$, from 70.2% to 66.0% on the Mix split. This trend suggests that while anchor is crucial for target identity, an excessively large coefficient may over-correct the latent embeddings, potentially overriding the spatial cues captured by the original segmentation token. Furthermore, the Null split performance was most stable at $\alpha = 0.1$, indicating that a carefully calibrated refinement better preserves the model's ability to suppress false-positive segments when no valid target is present. Consequently, we adopt $\alpha = 0.1$ as the default setting to maintain an optimal balance between identity anchoring and segmentation representation.

A.4.4 Ablation on mask decoder

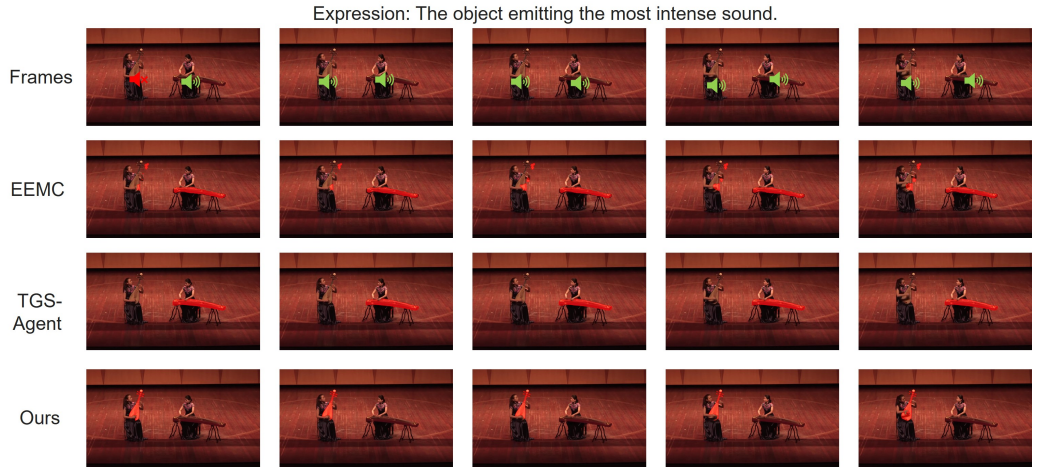
We further evaluated the impact of different mask decoders on the segmentation performance. Specifically, we replaced the specialized mask decoder in our SAIL with four baseline configurations: a standard Mask2Former (M2F) decoder [3], a vanilla Segment Anything Model (SAM) decoder [15], a query propagation (QP) enhanced SAM decoder as proposed in OISA-1B, and a SAM2 decoder [35] as adopted in SAM2-LOVE. As presented in Table 12, the transition from M2F to SAM yielded a noticeable improvement, demonstrating the superior zero-shot localization capabilities of the foundation decoder. The inclusion of QP, which enables frame-by-frame online updates of the query to capture temporal motion and contextual information, further boosted the performance. This highlights the importance of modeling the target's dynamic evolution across the video timeline. Similarly, the inclusion of SAM2 also achieved higher performance than static M2F or SAM, especially on the Unseen split. Notably, our proposed SAIL outperformed all the above configurations, achieving the highest scores across all metrics. While QP focuses on smooth motion tracking through query updates and SAM2 employs memory bank updates, SAIL's specialized mask decoder enables shift-aware prompting and memory-based propagation that provide a more robust mechanism for consistent segmentation, leading to 1.2% and 2.2% improvements in $\mathcal{J}\&\mathcal{F}$ on the Mix split, respectively. The results confirm that SAIL's specialized mask decoder effectively integrates shift-aware representations, enabling more robust and consistent segmentation in complex multimodal scenarios.

A.5 More qualitative results

In this section, we present more qualitative results in Fig. 4 and Fig. 5, illustrating the model's performance on the Seen and Unseen splits of Ref-AVSBench [45], respectively. For each example, we show video frames, expression, and predictions by different models. The visualized examples highlight SAIL's superior capability in reasoning over complex tri-modal inputs while maintaining temporal identity consistency. In Fig. 4(a), the scene contains two potential targets. While EEMC



(a)



(b)

Figure 4: Additional qualitative results from EEMC [45], TGS-Agent [57], and our proposed SAIL on the Seen split from the Ref-AVSBench [45] dataset.

failed to localize the focus of the expression, which was distracted by the visual prominence of the hair-dryer, and TGS-Agent mistakenly identified the hair-dryer as the primary sounding object, SAIL correctly segmented the dog. This success is attributed to our shift-aware refinement, which leverages the `[query]` token to filter out acoustic distractors and maintain focus on the linguistically referred target. Fig. 5(b) depicts a challenging out-of-distribution scenario involving a "gun" and an "earphone". Considering the presence of a salient yellow earphone and the explicit phrasing "after the man", TGS-Agent was misled by the visual distractor of the earphone. In contrast, SAIL generated consistent and accurate masks for the target throughout the timeline. This confirms that our identity-guided anchoring effectively generalizes to novel categories by prioritizing semantic logic over raw visual bias. We further evaluated SAIL on the R^2-AVSBench [57] to qualitatively assess its performance in scenarios requiring complex semantic reasoning. Fig. 6 presents two examples for implicit expressions. Even when the target is described through indirect functional or acoustic attributes, our method demonstrated remarkable robustness in preserving target identity. This visual evidence confirms that the shift-aware refinement mechanism effectively bridges the gap between linguistic reasoning and pixel-wise consistency, maintaining a stable mask even as the visual and auditory contexts evolve.

Despite its robust performance, SAIL encounters limitations in specific edge cases, as illustrated in Fig. 7. As shown in the top example of the figure, when a target is both relatively small and partially occluded (*e.g.*, the flute hold by the woman in the final frame), the refined segmentation tokens may lack sufficient spatial density to guide the mask decoder accurately. In such instances, the latent information is dominated by background features, leading to fragmented or missing masks.

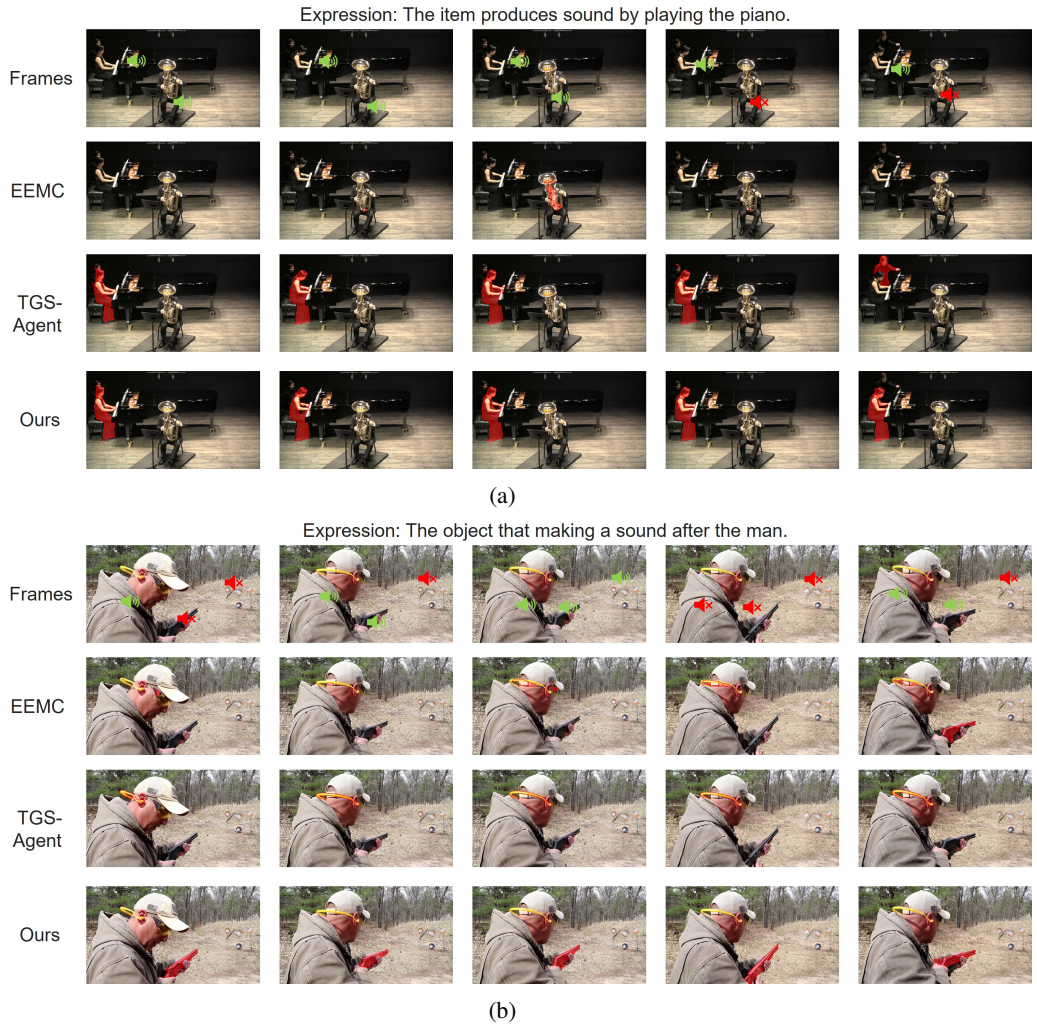


Figure 5: Additional qualitative results from EEMC [45], TGS-Agent [57], and our proposed SAIL on the Unseen split from the Ref-AVSBench [45] dataset.

In scenarios involving multiple overlapping sound sources paired with imprecise or underspecified linguistic expressions (*e.g.*, Fig. 7(b)), the model may suffer from target ambiguity. When the prompt does not sufficiently distinguish between two active sounding objects, the model may default to the most visually salient one, leading to an identity drift. This limitation is further exacerbated in tri-modal settings, where ambiguity in referring expressions, coupled with competing auditory and visual cues, makes consistent target grounding particularly challenging. These failure modes suggest that while our latent refinement significantly mitigates target shift, there remains room to improve the model's sensitivity to fine-grained spatial details and its ability to resolve high-level semantic contradictions. We leave the integration of multi-scale spatial feedback and interactive disambiguation for future work.

A.6 Complexity analysis

To evaluate computational efficiency, we compared SAIL with two representative baselines, SAM2-LOVE and Crab. All measurements were conducted on a single NVIDIA H100 GPU with a batch size of 1 for fair comparison. As shown in Table 13, SAIL incurs higher computational overhead than SAM2-LOVE and Crab, with an inference latency of 128 ms, throughput of 7.8 FPS, and peak memory usage of 22.5 GB. This additional cost mainly stems from

Method	Latency (ms)	FPS	Memory (GB)
SAM2-LOVE	54	18.5	12.4
Crab	115	8.7	18.2
SAIL	128	7.8	22.5

Table 13: Latency and memory usage.

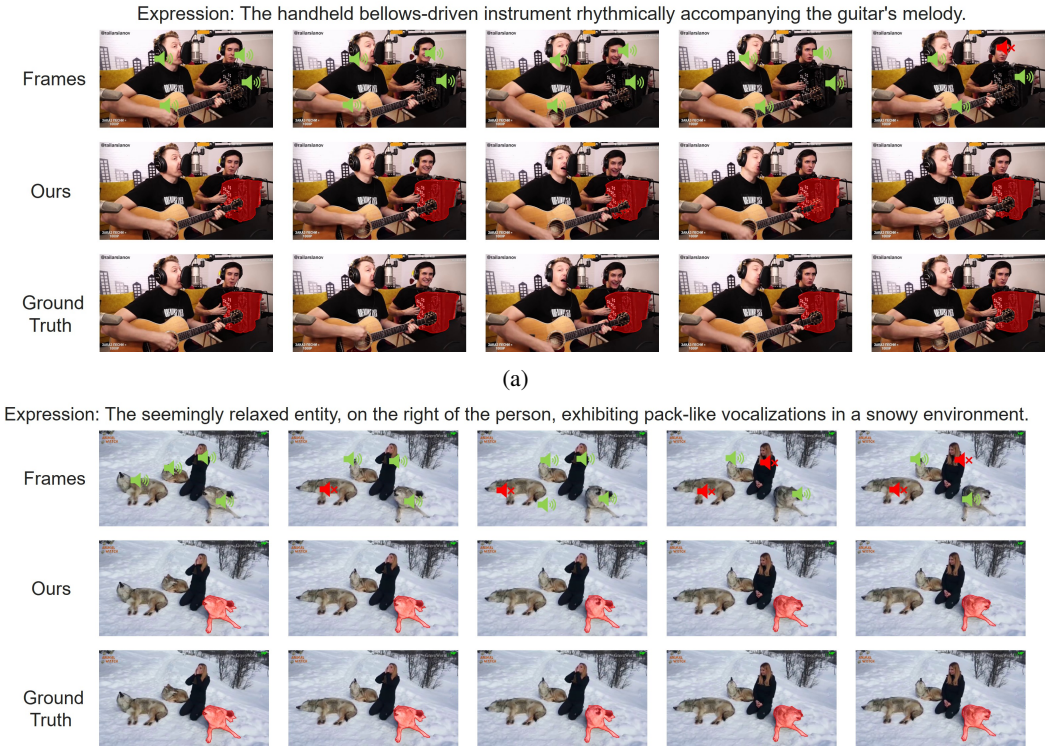


Figure 6: Qualitative results on samples from the R^2-AVSBench [57] dataset. The two examples are from the Seen and Unseen splits, respectively.

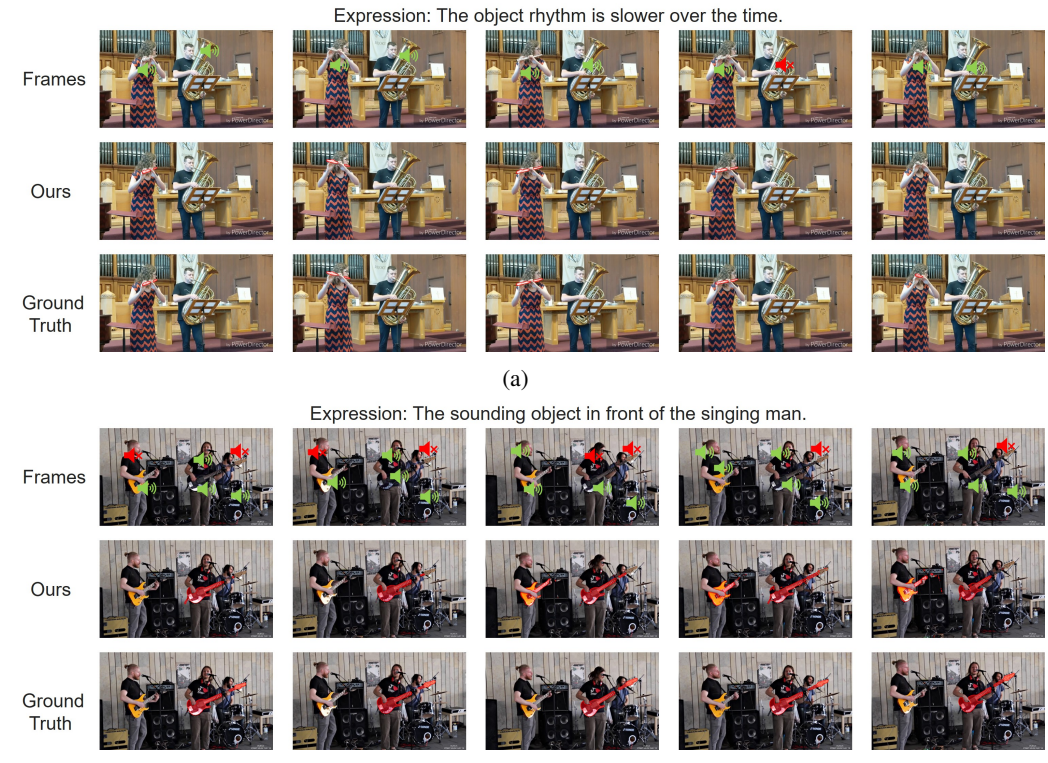


Figure 7: Failure cases for the proposed SAIL on the Ref-AVSBench [45] dataset.

the shift-aware refinement and identity-guided reasoning modules introduced to mitigate target shift. Compared with Crab, SAIL introduces only a moderate latency increase (+13 ms) while yielding substantial performance gains on challenging target-shift scenarios, achieving a +26.3% improvement in $\mathcal{J}\&\mathcal{F}$ on the Mix split in Table 8. Notably, SAIL is trained end-to-end and performs inference without post-processing, additional instruction-tuning data, or multi-stage pipelines with external components.

A.7 Discussion and future work

While SAIL demonstrates state-of-the-art performance in Ref-AVS, several avenues remain for further exploration and refinement. A primary limitation of our current framework is its reliance on per-frame token generation. While this approach ensures high precision for the 10-second clips typical of the Ref-AVSBench, it presents a computational bottleneck for long-form video analysis. Processing extended sequences requires more efficient temporal abstraction. Future work could draw inspiration from Ref-VOS techniques [7, 47], implementing key-frame selection and sparse-to-dense latent propagation. In addition, SAIL operates in a *one-shot* inference mode, where the model processes a static prompt and returns a set of masks. However, human-AI collaboration often requires a more dynamic and interactive referring process [19, 54]. We envision extending SAIL to support multi-turn dialogues where users can provide corrective feedback (*e.g.*, "No, the person on the left") or adjust the segmentation focus through iterative natural language refinement. Developing a more intelligent and controllable interface would allow SAIL to operate as a proactive assistant in complex video editing and processing contexts. Last but not least, fine-grained multi-object relational reasoning is a promising direction for Ref-AVS. Integrating scene-graph reasoning [32] or object-level relational [24] modules could strengthen the model's fine-grained spatial understanding.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract claims that the paper introduces a research problem and proposes a method which has key novelties in the model design and achieves state-of-the-art performance. The introduction clearly illustrates the contributions of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: It's in the Section A.5 of the supplementary materials.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The implementation details demonstrate the hyperparameters and training strategy. Supplementary materials also provide further details due to the page limit.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets are publicly available benchmarks. The code will be released in Open Access, under some license.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: They are explained in the method, experiment, and supplementary material sections.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: It would be too computationally expensive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We mention the information in the implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer:[Yes]

   Justification: The paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The models and data used in the paper are properly credited and the license and terms of use are properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: They will be provided upon acceptance of the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The proposed model leverages an open-source MLLM as the component of the framework. We introduce this in the method section.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.