RETHINKING AUDIO-VISUAL SEGMENTATION WITH ASYMMETRIC VISUAL ENTITY CONSTRUCTION AND AUDIO QUERYING

Anonymous authors
Paper under double-blind review

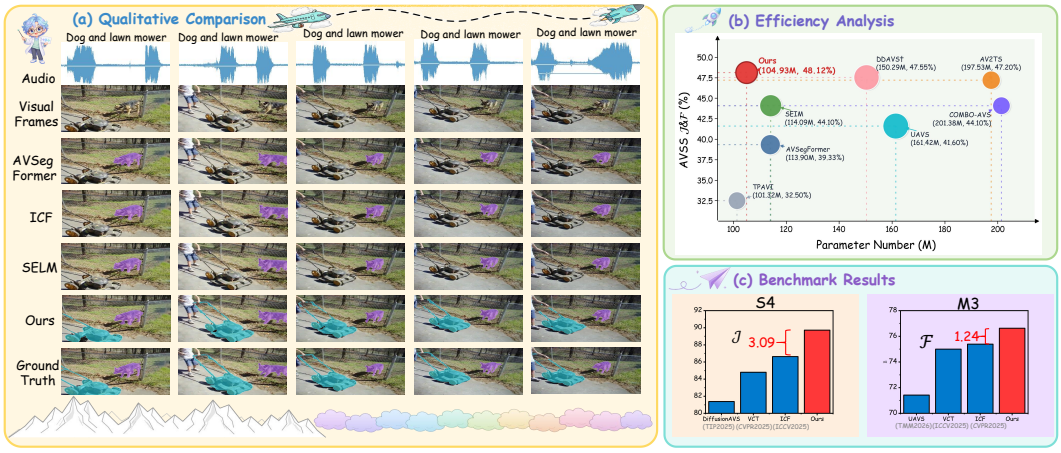

Figure 1: Qualitative and quantitative comparisons for audio-visual segmentation. (a) Our method provides more accurate and consistent sounding-object segmentation than existing approaches. (b) Efficiency comparison on AVSS in terms of parameters, FLOPs, and $\mathcal{J}\&\mathcal{F}$. (c) Performance comparisons on S4 and M3 using $\mathcal{J}$ and $\mathcal{F}$, respectively.

ABSTRACT

Audio-Visual Segmentation (AVS) requires associating unlocalized audio cues with spatially dense visual entities. Existing methods mostly rely on symmetric fusion, which entangles representations and induces spurious correlations. This overlooks a fundamental inductive bias: *vision delineates entity locations, whereas audio merely verifies their acoustic activation*. Motivated by this asymmetry, we reformulate AVS as an asymmetric entity-querying problem and propose AEQNet that separates visual entity construction from candidate-grounded acoustic verification through four complementary components. Firstly, we propose the Hierarchical Visual Candidate Constructor (HVCC), which leverages learnable hierarchical aggregation and mask-conditioned abstraction to construct compact object-centric candidates, addressing the spatial ambiguity of direct dense interactions. Secondly, we propose the Coarse-to-Fine Audio Query Router (CF-AQR), which employs coarse semantic matching and candidate-constrained fine verification to progressively determine candidate audio activation, resolving simultaneous multi-source ambiguities. Thirdly, we propose the Factorised Semantic Composer (FSC), which explicitly combines spatial membership, semantic identity, and audio activation for structured semantic predictions, avoiding entangled multimodal decoding. Finally, we propose Progressive Candidate-Grounded Counterfactual Query Learning (PCQL), which leverages progressive optimization and hard audio counterfactuals to encourage audio-sensitive candidate activations, thereby reducing visual shortcuts and spurious correlations. Extensive experiments on standard AVS benchmarks demonstrate the effectiveness and generality of our framework, which achieves strong and competitive performance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

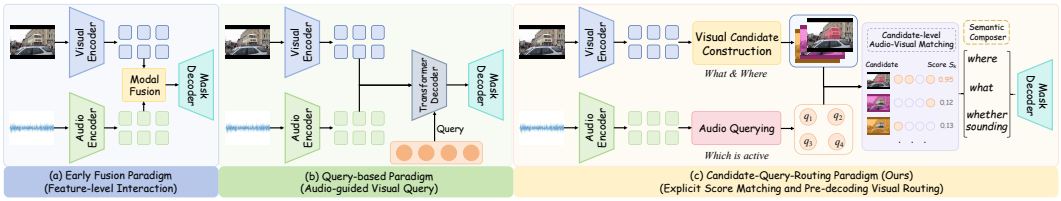

Figure 2: Comparison of AVS paradigms. Fusion- and query-based methods entangle audio-visual correspondence within feature interaction or decoding. In contrast, our candidate-query-routing paradigm explicitly constructs visual candidates and performs candidate-level acoustic matching.

1 INTRODUCTION

Audio-Visual Segmentation (AVS) aims to identify and segment sounding objects in videos. The core challenge is associating spatially unlocalized audio cues with visual entities to determine which objects are acoustically active, requiring robust cross-modal alignment for precise mask prediction. AVS has broad applications in embodied AI (Gupta et al., 2021), autonomous driving (Wang et al., 2024), video editing (Li et al., 2024a), and surveillance (Cristani et al., 2007). Fig. 1 compares our method with existing approaches in segmentation quality, efficiency, and benchmark performance.

Existing AVS architectures can be broadly organized into two dominant paradigms: **Feature-Fusion Encoder–Decoder** and **Audio-Centric Query Transformer**. The first paradigm (Yang et al., 2024; Liu et al., 2025; Gong et al., 2025; Zhu et al., 2025) separately extracts audio and visual representations and progressively integrates them through cross-modal fusion before mask decoding (see Fig. 2 (a)). While recent methods have advanced this paradigm by exploring bilateral cross-modal dependencies (Yang et al., 2024) or progressive semantic fusion (Liu et al., 2025), they persistently treat AVS as a symmetric problem. However, modality roles are inherently asymmetric: *vision determines what entities exist and where they are located, whereas audio primarily provides evidence about which visible entities are acoustically active*. Such dense cross-modal coupling inadvertently entangles spatial identity with acoustic state, rendering models vulnerable to spurious correlations and visual shortcuts. The second paradigm, Audio-Centric Query Transformers (Huang et al., 2023; Gao et al., 2024; Tian et al., 2026; Huang et al., 2025; Peng et al., 2026; Lv et al., 2026), employs latent tokens as queries to retrieve target acoustic cues from dense visual representations (see Fig. 2 (b)). Representative works utilize audio-conditioned queries (Huang et al., 2023; Gao et al., 2024), mixed-audio decomposition (Tian et al., 2026), or vision-derived iterative retrieval (Huang et al., 2025) for cross-modal alignment. Nevertheless, existing query-based architectures still largely follow the Transformer object-query paradigm, where queries simultaneously undertake object discovery, spatial localization, cross-modal matching, representation refinement, and mask prediction. Such query overloading is particularly problematic, as spatially unlocalized audio directly corrupts *what an entity is*, while repeated multimodal refinement re-entangles visual identity with *whether it is acoustically active*. Motivated by the object-centric representation principle (Brady et al., 2023), we fundamentally rethink AVS from an entity-centric perspective and propose a novel asymmetric entity-querying architecture (see Fig. 2 (c)). Rather than densely coupling modalities or employing audio queries for object discovery, our framework first constructs audio-independent visual candidates, restricting audio strictly to acoustic verification. By explicitly decoupling visual entity construction from acoustic querying, this formulation successfully circumvents both the representation entanglement of fusion models and the query overloading of audio-centric transformers.

In this work, we propose the Asymmetric Entity-Querying Network (AEQNet) to address representation entanglement and query overloading. Firstly, we introduce the Hierarchical Visual Candidate Constructor (HVCC) to alleviate spatial ambiguity from dense cross-modal interactions. HVCC aggregates multi-level visual features and performs mask-conditioned abstraction to build compact, object-centric candidates. Secondly, we develop the Coarse-to-Fine Audio Query Router (CF-AQR) to resolve ambiguous audio-to-entity correspondences. It estimates candidate acoustic activation by combining coarse semantic retrieval with local verification. Thirdly, we propose the Factorized Semantic Composer (FSC) to prevent modality re-entanglement during decoding. FSC factorizes candidate membership, visual identity, and acoustic activation to form a feature context, which is residually injected into dense visual representations. Finally, we introduce Progressive Candidate-Grounded Counterfactual Query Learning (PCQL) to suppress visual shortcuts. PCQL re-queries visual candidates with hard counterfactual audio to enforce genuinely audio-dependent activations.

Collectively, this paradigm of explicit entity construction and acoustic querying effectively supersedes conventional dense fusion. At the same time, we demonstrate the effectiveness of our method through extensive experiments. Our contributions are summarized as follows:

- We propose AEQNet, an asymmetric entity-querying framework that decouples visual entity construction from acoustic verification for audio-visual segmentation. Specifically, we design the HVCC to adaptively aggregate different visual features and construct compact object-centric candidates, reducing the spatial ambiguity of dense audio-visual interaction.
- We propose a CF-AQR that combines coarse semantic matching with candidate-constrained fine verification to establish reliable audio-to-entity correspondence under ambiguous and multi-source audio.
- We propose an FSC that factorizes candidate membership, audio activation, and visual identity, and residually injects the resulting candidate context into dense visual representations for semantic prediction.
- We propose PCQL that contrasts positive and hard counterfactual audio over the same visual candidates to suppress visual shortcuts and enforce genuinely audio-dependent decisions. Extensive experiments on multiple AVS benchmarks demonstrate the effectiveness and generality of our proposed architecture.

2 Related Work

2.1 Audio Visual Segmentation

Audio-visual segmentation requires accurately delineating sounding entities dynamically driven by acoustic cues. Existing architectures predominantly rely on dense cross-modal feature fusion (e.g., TPAVI (Zhou et al., 2022), SELM (Li et al., 2024a)), latent query interactions (e.g., DDESeg (Liu et al., 2025)), or auxiliary frameworks (e.g., ICF (Zha et al., 2025)). However, fusion-based methods often entangle multimodal representations, query-centric methods overload latent queries with conflicting spatial and semantic tasks, and auxiliary models inherently increase architectural complexity. To address these limitations, we propose a novel asymmetric entity-querying paradigm. By explicitly constructing audio-independent visual entities and using the audio modality solely for candidate-level verification, our method cleanly separates visual entity construction from acoustic grounding without relying on external auxiliary models.

2.2 Multimodal Query Learning

Query-based learning employs latent tokens for object-centric abstraction and selective cross-modal interaction (Jaegle et al., 2021; Locatello et al., 2020; Li et al., 2023). While recent audio-visual segmentation (AVS) frameworks (Gao et al., 2024; Huang et al., 2025) heavily rely on such modality-derived queries, they risk query collapse (Shah et al., 2024) and persistently suffer from query overloading by forcing tokens to simultaneously manage object discovery, alignment, and mask prediction. To address this structural bottleneck, our approach explicitly decouples entity representation from cross-modal querying. We construct robust visual entities independently, restricting audio queries solely to candidate-level acoustic verification.

3 Methodology

As illustrated in Figure 3, given an audio-visual clip consisting of visual input $\mathbf{V} \in \mathbb{R}^{B\times T\times 3\times H_0\times W_0}$ and audio input $\mathbf{A} \in \mathbb{R}^{B\times T\times L_a}$ (where B and T denote the batch size and temporal length), the visual encoder first extracts dense hierarchical features $\{\mathbf{F}_i^v\}_{i=1}^{L}$. Concurrently, the audio encoder produces global acoustic features $\mathbf{F}^a \in \mathbb{R}^{B\times T\times d_a}$. Next, the HVCC component distills the dense visual hierarchy into $K = 8$ explicit, object-centric candidates, yielding spatial candidate masks $\mathbf{M} \in \mathbb{R}^{B\times T\times K\times h\times w}$ and compact embeddings $\mathbf{Z} \in \mathbb{R}^{B\times T\times K\times d}$. CF-AQR then utilizes the audio cues to query these visual candidates, computing candidate-level acoustic support scores $\mathbf{S} \in \mathbb{R}^{B\times T\times K}$. Subsequently, FSC integrates the normalized candidate spatial memberships, acoustic support scores, and candidate embeddings to form a comprehensive candidate-conditioned

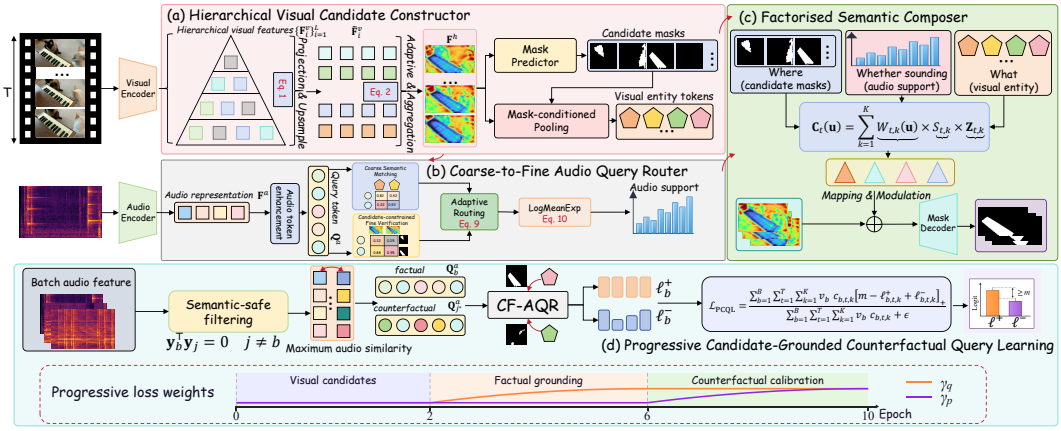

Figure 3: Overview of AEQNet. Visual and audio features are extracted using separate encoders. HVCC constructs audio-independent candidate masks and entity tokens from hierarchical visual features. CF-AQR estimates candidate-level audio support through coarse semantic matching and candidate-constrained fine verification. FSC composes candidate membership, audio support, and visual identity into a context that is residually injected into visual features for mask decoding. During training, PCQL contrasts factual and hard counterfactual audio queries over the same candidates with progressive loss weighting to suppress visual shortcuts.

context. This context is residually injected back into the dense visual representation to decode the final segmentation masks. During training, PCQL rigorously calibrates the audio-candidate correspondences via paired factual and counterfactual queries. Overall, our framework shifts dense multimodal interaction toward explicit candidate-level reasoning for robust segmentation.

3.1 Hierarchical Visual Candidate Constructor

Dense visual representations preserve rich spatial details but intrinsically lack the explicit entity structure required for audio-conditioned reasoning. To establish this structure independently of audio evidence, we introduce the HVCC to distill hierarchical visual representations into a compact set of explicit visual candidates. Given the hierarchical visual features $\{\mathbf{F}_i^v\}_{i=1}^L$, we first project each feature into a shared d-dimensional space:

$$\widetilde{\mathbf{F}}_i^v = Up_i(Conv_i^{1\times1}(\mathbf{F}_i^v)), \tag{1}$$

where $Conv_i^{1\times1}(\cdot)$ and $Up_i(\cdot)$ denote the 1×1 convolution operation and the spatial upsampling operation, respectively. To balance complementary information across feature hierarchies, we assign each level a learnable scalar ω_i and normalize these scalars into fusion coefficients α_i. The weighted features are then refined to obtain the hierarchical representation $\mathbf{F}^h$:

$$\alpha_i = \frac{\exp(\omega_i)}{\sum_{j=1}^L \exp(\omega_j)}, \qquad \mathbf{F}^h = Conv^{3\times3}(\delta(GN(Conv^{3\times3}(\sum_{i=1}^L \alpha_i \widetilde{\mathbf{F}}_i^v)))), \tag{2}$$

where $Conv^{3\times3}(\cdot)$ represents the 3×3 convolution operation, while $GN(\cdot)$ and $\delta(\cdot)$ refer to the Group Normalization and GELU activation functions, respectively. Conditioned on this fused representation, we instantiate explicit visual candidates by decoding their spatial masks $\mathbf{M}$ and distilling candidate-specific embeddings $\mathbf{Z}_{t,k}$:

$$\mathbf{M} = \sigma\big(\mathcal{G}_m(\mathbf{F}^h)\big), \qquad \mathbf{Z}_{t,k} = \frac{\sum_x \mathbf{M}_{t,k}(x)\mathbf{F}_t^h(x)}{\sum_x \mathbf{M}_{t,k}(x) + \epsilon}, \tag{3}$$

where $\mathcal{G}_m(\cdot)$ denotes the candidate mask predictor, while $\sigma(\cdot)$ and x refer to the sigmoid function and spatial location indices, respectively, with ϵ serving as a small constant for numerical stability. Specifically, $\mathcal{G}_m(\cdot)$ comprises the MLP projections, spatial upsampling operations, and feature concatenation, followed sequentially by a 1×1 convolution operation for feature fusion, dropout, and a final 1×1 convolution prediction layer. During training, each frame $(b,t)\in\mathcal{I}$ has ground-truth masks $\mathbf{G}_{b,t,r}$, where $\mathcal{I}$ denotes the local mini-batch and r indexes the foreground target masks within

each frame. Here, r indexes target masks, not necessarily individual sounding objects. "Object-centric" refers to the mask-conditioned representation of each candidate. Gradient-free Hungarian matching produces a set $\mathcal{P}$ of matched tuples (b, t, k, r), pairing candidate k with target mask r. The candidate supervision loss $\mathcal{L}_{\text{cand}}$ is

$$\mathcal{L}_{\text{cand}} = \frac{1}{\max(|\mathcal{P}|, 1)} \sum_{(b,t,k,r)\in\mathcal{P}} [\mathcal{L}_{\text{ce}}(\mathbf{U}_{b,t,k}, \mathbf{G}_{b,t,r}) + \mathcal{L}_{\text{dice}}(\mathbf{M}_{b,t,k}, \mathbf{G}_{b,t,r}) + \mathcal{E}_{b,t,k,r}], \tag{4}$$

where $\mathbf{U}$ represents the predicted mask logits. The objective function comprises the mean pixel-wise binary cross-entropy ($\mathcal{L}_{\text{ce}}$), Dice loss ($\mathcal{L}_{\text{dice}}$), and a classification loss ($\mathcal{E}$). Specifically, $\mathcal{E}$ is formulated as a foreground cross-entropy loss, defaulting to zero in classifier-free configurations. Ground-truth labels are re-indexed by omitting the background, designating class 0 for binary targets. Furthermore, the loss is exclusively computed over matched candidates, such that an empty assignment set $\mathcal{P}$ yields a zero penalty.

3.2 Coarse-to-Fine Audio Query Router

Although HVCC exposes potential visual entities, visual existence alone cannot determine whether an entity is supported by the corresponding audio. We therefore propose the CF-AQR, which progressively evaluates each visual candidate through global semantic matching and local spatial verification. We first transform the aligned audio representation $\mathbf{F}^a$ into Q audio query tokens for each sample at every time step. Each token is a d_q-dimensional vector for candidate-level acoustic matching, forming $\mathbf{Q}^a \in \mathbb{R}^{B\times T\times Q\times d_q}$:

$$\mathbf{Q}^a = \mathcal{H}_a(\mathbf{F}^a), \tag{5}$$

where $\mathcal{H}_a(\cdot)$ denotes the audio token enhancement operation (Liu et al., 2025), $Q = 5$ specifies the number of queries per time step, and $d_q = d$ matches the visual candidate embedding dimension. Before audio querying, we construct detached visual references:

$$\mathbf{M}^q = \text{sg}(\mathbf{M}), \qquad \mathbf{Z}^q = \text{sg}(\mathbf{Z}), \qquad \mathbf{F}^{h,q} = \text{sg}(\mathbf{F}^h), \tag{6}$$

where $\text{sg}(\cdot)$ preserves the forward value while blocking backward gradients. Both coarse matching and fine verification operate exclusively on these detached visual references. Hereafter, $F^{h,q}$ is treated as an implicit shared input to CF-AQR and omitted from its argument list for notational brevity. We then compute the coarse semantic relation $r^c_{t,q,k}$ between the q-th audio query and the k-th visual candidate:

$$r^c_{t,q,k} = \frac{(\overline{\mathbf{Q}}^a_{t,q})^\top \overline{\mathbf{Z}}^q_{t,k}}{\tau_c}, \qquad \overline{\mathbf{Q}}^a_{t,q} = \frac{\mathbf{Q}^a_{t,q}}{\max(\|\mathbf{Q}^a_{t,q}\|_2, \epsilon)}, \qquad \overline{\mathbf{Z}}^q_{t,k} = \frac{\mathbf{Z}^q_{t,k}}{\max(\|\mathbf{Z}^q_{t,k}\|_2, \epsilon)}, \tag{7}$$

where τ_c denotes the temperature coefficient for coarse audio-candidate matching and is set to 0.07. To verify spatial consistency, we compute a fine-grained relation $r^f_{t,q,k}$ inside each candidate region:

$$r^f_{t,q,k} = \frac{\sum_x \mathbf{M}^q_{t,k}(x)\, \text{sim}(\psi_a(\mathbf{Q}^a_{t,q}), \psi_v(\mathbf{F}^{h,q}_t(x)))}{\sum_x \mathbf{M}^q_{t,k}(x) + \epsilon}, \tag{8}$$

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity function, while $\psi_a(\cdot)$ and $\psi_v(\cdot)$ represent an audio projection consisting of linear layers equipped with Layer Normalization and GELU activation, and a 1×1 convolution operation projecting visual features into the shared query space, respectively. These two relations are adaptively integrated through a routing mechanism balancing semantic compatibility and spatial verification:

$$g_{t,q,k} = \sigma(\mathcal{G}_r(r^c_{t,q,k}, r^f_{t,q,k})), \quad r_{t,q,k} = g_{t,q,k} r^c_{t,q,k} + (1 - g_{t,q,k}) r^f_{t,q,k}, \tag{9}$$

where $\mathcal{G}_r(\cdot)$ denotes a routing gate composed of linear layers with GELU activation. We further aggregate the Q query responses using a query-count-invariant LogMeanExp operator (Pinheiro & Collobert, 2015), transforming the aggregated response $\ell_{t,k}$ into the candidate support score $S_{t,k}$:

$$\ell_{t,k} = \tau_p[\log \sum_{q=1}^{Q} \exp(\frac{r_{t,q,k}}{\tau_p}) - \log Q], \qquad S_{t,k} = \sigma(\ell_{t,k}), \tag{10}$$

where τ_p denotes the pooling temperature and is set to 0.5. Consequently, the factual and counterfactual query objectives update the trainable query-branch parameters but do not back-propagate into the network through the visual references. We supervise factual audio grounding using binary cross-entropy over all candidates:

$$\mathcal{L}_{\text{query}} = \frac{1}{K|\mathcal{I}|} \sum_{(b,t)\in\mathcal{I}} \sum_{k=1}^{K} [\text{softplus}(\ell_{b,t,k}) - o_{b,t,k}\ell_{b,t,k}], \tag{11}$$

where $\text{softplus}(\cdot)$ denotes the Softplus function (Dugas et al., 2000), and $o_{b,t,k} = \mathbb{I}[\exists r : (b,t,k,r) \in \mathcal{P}]$, with $\mathcal{P}$ obtained by Hungarian matching (Kuhn, 1955).

3.3 Factorised Semantic Composer

To circumvent the reintroduction of entangled representations caused by dense cross-modal fusion, we propose the FSC, which explicitly decomposes candidate-level reasoning into three complementary factors: candidate membership, audio support, and visual identity. Given the candidate masks $\mathbf{M}$, candidate embeddings $\mathbf{Z}$, and candidate support scores $\mathbf{S}$, we first derive a null-aware spatial membership for each visual candidate:

$$W_{t,k}(\mathbf{u}) = \frac{\mathbf{M}_{t,k}(\mathbf{u})}{1 + \sum_{j=1}^{K} \mathbf{M}_{t,j}(\mathbf{u})}, \tag{12}$$

where $\mathbf{u}$ denotes a spatial position and $W_{t,k}(\mathbf{u})$ represents the normalized membership of the k-th candidate at temporal step t. The constant 1 reserves an implicit null or background mass that prevents every spatial position from being forcibly assigned to one of the instantiated candidates. We then explicitly compose candidate membership, audio support, and visual identity:

$$\mathbf{C}_t(\mathbf{u}) = \sum_{k=1}^{K} \underbrace{W_{t,k}(\mathbf{u})}_{\text{candidate membership}} \times \underbrace{S_{t,k}}_{\text{audio support}} \times \underbrace{\mathbf{Z}_{t,k}}_{\text{visual identity}}. \tag{13}$$

Accordingly, the composed candidate context is collected as $\mathbf{C} \in \mathbb{R}^{B\times T\times d\times h\times w}$, representing where a candidate appears, whether it is acoustically active, and what visual entity it describes. Rather than directly replacing the dense visual representation, we project the composed candidate context and residually inject it into the hierarchical visual feature $\mathbf{F}_t^o$:

$$\mathbf{F}_t^o = \mathbf{F}_t^h + \gamma_f(e)\, Conv^{1\times1}(\mathbf{C}_t), \qquad \mathbf{Y}_t^l = Conv^{3\times3}(\mathbf{F}_t^o), \tag{14}$$

where $\gamma_f(e) \in [0,1]$ is a dynamic scheduling coefficient dependent on the training epoch e, which progressively regulates the influence of candidate-level audio evidence during the optimization process. Finally, $\mathbf{Y}^l$ represents the intermediate, low-resolution semantic prediction map, which is subsequently bilinearly upsampled to yield the final output $\widehat{\mathbf{Y}}$. The visual residual preserves spatial detail but leaves a possible visual shortcut. PCQL encourages audio-sensitive candidate scores without guaranteeing audio-dependent final predictions.

3.4 Progressive Candidate-Grounded Counterfactual Query Learning

Although CF-AQR evaluates visual candidates using audio cues, it risks overfitting to visual co-occurrence priors, assigning spuriously high responses to salient entities regardless of actual audio consistency. To mitigate this bias, we propose PCQL, which explicitly calibrates the cross-modal correspondence by contrasting *factual* and *counterfactual* audio queries against identical visual candidates. For the b-th sample, we construct a semantic-safe negative set to prevent false negatives and select the hardest valid audio sample j^*:

$$j^* = \arg\max_{j\in\mathcal{N}_b} \text{sim}\big(\mathbf{F}_b^a, \mathbf{F}_j^a\big), \qquad \text{where} \quad \mathcal{N}_b = \{j \neq b \mid \mathbf{y}_b^\top \mathbf{y}_j = 0\}, \tag{15}$$

where $\mathbf{y}_b$ denotes the foreground semantic-presence vector of sample b, constructed from the dataset-provided video-level category metadata. We subsequently query the fixed visual candidates using both the factual audio tokens $\mathbf{Q}_b^a$ and the selected counterfactual audio tokens $\mathbf{Q}_{j^*}^a$:

$$\boldsymbol{\ell}_b^+ = \text{CF-AQR}(\mathbf{Q}_b^a, \mathbf{M}_b^q, \mathbf{Z}_b^q), \qquad \boldsymbol{\ell}_b^- = \text{CF-AQR}\big(\mathbf{Q}_{j^*}^a, \mathbf{M}_b^q, \mathbf{Z}_b^q\big). \tag{16}$$

Table 1: Quantitative comparison of $\mathcal{J}$, $\mathcal{F}$ and $\mathcal{J}\&\mathcal{F}$ on the S4 (Zhou et al., 2022), M3 (Zhou et al., 2022), and AVSS (Zhou et al., 2025) datasets. ↑ indicates that higher values are better. Best performance in **bold**, second best underlined. ‡ denotes unavailable data.

Method	Source	Backbone	AVS-Object (S4)			AVS-Object (M3)			AVSS		
			$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
AVSegFormer (Gao et al., 2024)	AAAI24	ResNet-50	80.67	76.54	84.80	56.17	49.53	62.80	27.12	24.93	29.30
BAVS (Liu et al., 2024a)	TMM24	ResNet-50	81.63	77.96	85.29	56.30	50.23	62.37	27.16	24.68	29.63
QDFormer (Li et al., 2024b)	CVPR24	ResNet-50	81.80	77.60	86.00	61.55	59.60	63.50	‡	‡	‡
UFE (Liu et al., 2024b)	CVPR24	ResNet-50	83.23	78.96	87.50	60.19	55.88	64.50	‡	‡	‡
CAVP (Chen et al., 2024a)	CVPR24	ResNet-50	83.84	78.78	88.89	61.48	55.82	67.14	32.83	30.37	35.29
SELM (Li et al., 2024a)	ACM MM24	ResNet-50	81.40	76.60	86.20	60.05	54.50	65.60	34.55	31.90	37.20
COMBO (Yang et al., 2024)	CVPR24	ResNet-50	85.90	81.70	90.10	60.55	54.50	66.60	35.30	33.30	37.30
CPM (Chen et al., 2024b)	ECCV24	ResNet-50	85.92	81.37	90.47	65.40	59.80	71.00	37.05	34.53	39.57
CQFormer (Lv et al., 2025)	TIP25	ResNet-50	81.05	75.45	86.64	63.56	54.99	72.12	28.38	26.33	30.42
UAVS (Lv et al., 2026)	TMM26	ResNet-50	81.01	75.71	86.30	61.62	53.42	69.81	28.74	26.57	30.90
Ours	-	ResNet-50	**89.09**	**86.70**	**91.47**	**69.06**	**64.92**	**73.19**	**38.06**	**35.15**	**40.96**
AVSBG (Hao et al., 2024)	AAAI24	PVT-v2	86.06	81.71	90.40	60.95	55.10	66.80	‡	‡	‡
AVSegFormer (Gao et al., 2024)	AAAI24	PVT-v2	85.98	82.06	89.90	63.83	58.36	69.30	39.33	36.66	42.00
UFE (Liu et al., 2024b)	CVPR24	PVT-v2	86.78	83.15	90.40	66.43	61.95	70.90	‡	‡	‡
SELM (Li et al., 2024a)	ACM MM24	PVT-v2	87.35	83.50	91.20	65.80	60.30	71.30	44.10	41.30	46.90
COMBO (Yang et al., 2024)	CVPR24	PVT-v2	88.30	84.70	91.90	65.20	59.20	71.20	44.10	42.10	46.10
ICF (Zha et al., 2025)	ICCV25	PVT-v2	90.07	86.63	93.51	69.89	64.38	75.39	**48.16**	45.03	**51.28**
DiffusionAVS (Mao et al., 2025)	TIP25	PVT-v2	85.79	81.38	90.20	64.54	58.18	70.90	‡	‡	‡
VCT (Huang et al., 2025)	CVPR25	PVT-v2	88.45	84.80	92.10	68.50	62.00	75.00	47.10	44.70	49.50
CQFormer (Lv et al., 2025)	TIP25	PVT-v2	87.40	83.62	91.18	66.83	60.99	72.67	40.55	38.09	43.01
DDAVS (Tian et al., 2026)	ECCV26	PVT-v2	91.35	89.36	93.34	**73.60**	68.77	**78.43**	47.55	44.11	50.98
UAVS (Lv et al., 2026)	TMM26	PVT-v2	87.49	83.53	91.44	65.52	59.62	71.42	41.60	39.08	44.11
PGCNet (Zhou et al., 2026)	TMM26	PVT-v2	87.00	83.10	90.90	65.61	60.02	71.20	‡	‡	‡
AV2TS (Shen et al., 2026)	TMM26	PVT-v2	88.80	85.10	92.50	73.10	68.00	78.20	47.20	45.00	49.40
Ours	-	PVT-v2	**91.63**	**89.72**	**93.54**	72.98	**69.34**	76.63	48.12	**45.15**	51.08

Since both querying paths operate on identical visual candidates, the response discrepancy is induced strictly by the variation in audio evidence. To enforce discriminative audio-visual grounding, we apply a margin loss on the valid matched candidates:

$$\mathcal{L}_{\mathrm{PCQL}} = \frac{\sum_{b=1}^{B}\sum_{t=1}^{T}\sum_{k=1}^{K} v_b\, c_{b,t,k}[m - \ell^{+}_{b,t,k} + \ell^{-}_{b,t,k}]_{+}}{\sum_{b=1}^{B}\sum_{t=1}^{T}\sum_{k=1}^{K} v_b\, c_{b,t,k} + \epsilon}, \tag{17}$$

where $[z]_{+} = \max(z, 0)$, $c_{b,t,k} \in \{0, 1\}$ indicates whether the k-th candidate is matched with a sounding object, and $v_b \in \{0, 1\}$ indicates whether a valid semantic-safe counterfactual audio exists for the b-th sample. Furthermore, the margin m, set to 0.2, encourages the factual logit $\ell^{+}_{b,t,k}$ to exceed its counterfactual counterpart $\ell^{-}_{b,t,k}$ by at least m. The PCQL encourages relative logit separation but does not guarantee low support for mismatched audio.

3.5 Loss Function

We supervise the final prediction $\widehat{\mathbf{Y}}$ using a hybrid objective integrating pixel-wise and region-level constraints:

$$\mathcal{L}_{\mathrm{seg}} = \lambda_{\mathrm{ce}}\mathcal{L}_{\mathrm{ce}} + \lambda_{\mathrm{f}}\mathcal{L}_{\mathrm{focal}} + \lambda_{\mathrm{d}}\mathcal{L}_{\mathrm{dice}} + \lambda_{\mathrm{i}}\mathcal{L}_{\mathrm{iou}}, \tag{18}$$

where $\mathcal{L}_{\mathrm{ce}}$, $\mathcal{L}_{\mathrm{focal}}$, $\mathcal{L}_{\mathrm{dice}}$, and $\mathcal{L}_{\mathrm{iou}}$ denote complementary segmentation losses, weighted by $\lambda_{\mathrm{ce}} = 5$, $\lambda_{\mathrm{f}} = 5$, $\lambda_{\mathrm{d}} = 5$, and $\lambda_{\mathrm{i}} = 2$, respectively. The overall training objective combines $\mathcal{L}_{\mathrm{seg}}$ with candidate construction ($\mathcal{L}_{\mathrm{cand}}$), factual audio grounding ($\mathcal{L}_{\mathrm{query}}$), and counterfactual calibration ($\mathcal{L}_{\mathrm{PCQL}}$):

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{seg}} + \lambda_c\mathcal{L}_{\mathrm{cand}} + \gamma_q(e)\lambda_q\mathcal{L}_{\mathrm{query}} + \gamma_p(e)\lambda_p\mathcal{L}_{\mathrm{PCQL}}, \tag{19}$$

where $\lambda_c = 0.5$, $\lambda_q = 0.5$, and $\lambda_p = 0.1$ represent their respective balancing coefficients. To prevent unreliable audio supervision from disturbing immature visual candidates, $\gamma_q(e)$ and $\gamma_p(e)$ are progressively activated. Specifically, $\gamma_q(e)$ linearly increases from zero to one over $\Delta_q = 4$ epochs starting at epoch $e_q = 2$. Subsequently, $\gamma_p(e)$ follows an identical ramp schedule starting at epoch $e_p = 6$. We set $\gamma_f(e) = \gamma_p(e)$, so that FSC context injection follows the same starting epoch and linear ramp as counterfactual supervision. This explicitly staged optimization sequentially establishes visual candidates, factual audio grounding, and counterfactual calibration.

4 Experiments

4.1 Experimental Setup

Implementation details. We implement our model in PyTorch and train it on eight RTX 4090 GPUs, resizing all input frames to 224×224. Models are trained for 150 epochs using the AdamW

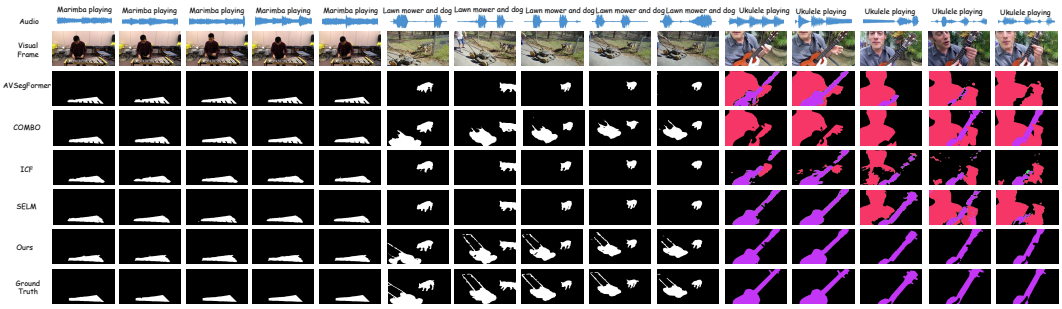

Figure 4: Qualitative comparison on S4, M3, and AVSS. From left to right: fine-grained object segmentation, multi-source audio-visual ambiguity, and semantic confusion. Our method produces more complete and accurate sounding-object masks.

Table 2: Ablation of the proposed components.

HVCC	CF-AQR	FSC	PCQL	S4 $\mathcal{J}\&\mathcal{F}$	S4 $\mathcal{J}$	S4 $\mathcal{F}$	M3 $\mathcal{J}\&\mathcal{F}$	M3 $\mathcal{J}$	M3 $\mathcal{F}$
				86.17	83.54	88.79	66.45	62.15	70.75
✓				88.35	85.70	90.99	67.98	63.60	72.36
✓	✓			89.87	87.74	92.00	69.97	66.00	73.93
✓	✓	✓		90.26	88.16	92.35	71.64	67.80	75.47
✓	✓	✓	✓	**91.63**	**89.72**	**93.54**	**72.98**	**69.34**	**76.63**

Table 3: Ablation of HVCC components.

Variant	S4 $\mathcal{J}\&\mathcal{F}$	S4 $\mathcal{J}$	S4 $\mathcal{F}$	M3 $\mathcal{J}\&\mathcal{F}$	M3 $\mathcal{J}$	M3 $\mathcal{F}$
LW→UA	91.03	88.97	93.08	71.97	68.33	75.60
SM→HM	90.87	88.85	92.88	72.09	68.12	76.06
w/o FR	90.82	88.58	92.79	72.67	68.73	76.61
w/o $\mathcal{L}_{\text{cand}}$	91.18	89.10	93.25	72.25	68.46	76.03
Ours	**91.63**	**89.72**	**93.54**	**72.98**	**69.34**	**76.63**

Table 4: Ablation studies of CF-AQR, FSC, and PCQL on S4 and M3.

CF-AQR Variant	S4 $\mathcal{J}\&\mathcal{F}$	S4 $\mathcal{J}$	S4 $\mathcal{F}$	M3 $\mathcal{J}\&\mathcal{F}$	M3 $\mathcal{J}$	M3 $\mathcal{F}$
w/o CM	90.80	88.76	92.83	72.49	68.64	76.34
w/o FV	90.88	88.85	92.91	72.68	68.73	76.63
w/o RG	91.22	89.14	93.30	72.13	68.50	75.75
AR→LS	90.93	88.85	93.00	72.90	69.09	76.70
SR→HR	91.19	89.07	93.31	72.80	68.96	76.63
Ours	**91.63**	**89.72**	**93.54**	**72.98**	**69.34**	**76.63**

FSC Variant	S4 $\mathcal{J}\&\mathcal{F}$	S4 $\mathcal{J}$	S4 $\mathcal{F}$	M3 $\mathcal{J}\&\mathcal{F}$	M3 $\mathcal{J}$	M3 $\mathcal{F}$
FC→DF	91.15	89.10	93.20	72.83	68.90	76.76
w/o AS	91.12	89.13	93.10	72.80	69.02	76.58
RI→DR	90.69	88.68	92.69	72.35	68.52	76.17
NM→SN	90.45	88.32	92.58	72.11	68.32	75.89
SG→E2E	90.75	88.69	92.81	72.55	68.89	76.21
Ours	**91.63**	**89.72**	**93.54**	**72.98**	**69.34**	**76.63**

PCQL Variant	S4 $\mathcal{J}\&\mathcal{F}$	S4 $\mathcal{J}$	S4 $\mathcal{F}$	M3 $\mathcal{J}\&\mathcal{F}$	M3 $\mathcal{J}$	M3 $\mathcal{F}$
SHN→SRN	91.00	88.95	93.04	72.57	68.74	76.40
w/o SSF	90.61	88.44	92.78	72.18	68.37	75.99
PO→JO	90.49	88.29	92.69	72.93	69.14	76.71
LM→PM	90.73	88.56	92.90	72.33	68.60	76.05
w/o $\mathcal{L}_{\text{PCQL}}$	90.26	88.16	92.35	71.64	67.80	75.47
Ours	**91.63**	**89.72**	**93.54**	**72.98**	**69.34**	**76.63**

optimizer (batch size 32, learning rate 1×10^{-4}). No additional post-processing is applied during inference. For fair comparison (Zha et al., 2025; Huang et al., 2025), we use ImageNet-pretrained ResNet-50 (He et al., 2016) and PVT-v2 (Wang et al., 2022) as visual backbones, alongside YouTube-8M-pretrained VGGish (Hershey et al., 2017) for audio feature extraction.

Datasets and Metrics. We evaluate our method on the widely used AVS-Object (Zhou et al., 2022) and AVSS (Zhou et al., 2025) benchmarks. AVS-Object contains 5-frame clips with class-agnostic binary annotations and is divided into a single-source S4 subset (3,452/740/740 for train/val/test) and a multi-source M3 subset (296/64/64). AVSS provides explicit semantic labels across 70 categories, comprising a total of 12,356 10-frame videos (8,498/1,304/1,554 splits). For comprehensive quantitative assessment, we adopt the Jaccard index $\mathcal{J}$, the F-score $\mathcal{F}$, and their mean $\mathcal{J}\&\mathcal{F}$, where higher values indicate consistently superior overall segmentation performance.

4.2 Comparison with the State-of-the-Art

We compare our method with recent AVS approaches: AVSegFormer (Gao et al., 2024), BAVS (Liu et al., 2024a), QDFormer (Li et al., 2024b), UFE (Liu et al., 2024b), CAVP (Chen et al., 2024a), SELM (Li et al., 2024a), COMBO (Yang et al., 2024), CPM (Chen et al., 2024b), CQFormer (Lv et al., 2025), ICF (Zha et al., 2025), DiffusionAVS (Mao et al., 2025), VCT (Huang et al., 2025), DDAVS (Tian et al., 2026), UAVS (Lv et al., 2026), PGCNet (Zhou et al., 2026), and AV2TS (Shen et al., 2026). Competitors follow their original protocols. We report results on the S4, M3, and AVSS benchmarks using ResNet-50 and PVT-v2 backbones.

Quantitative analysis. As shown in Table 1, our method establishes state-of-the-art results across most evaluation metrics with the ResNet-50 backbone, outperforming CPM by 3.17 and 3.66 $\mathcal{J}\&\mathcal{F}$ points on the S4 and M3 settings, respectively. The substantial improvements in $\mathcal{J}$ scores (e.g., +5.12 over CPM and +10.42 over COMBO on M3) further highlight our superior spatial localization capability for sounding objects. When equipped with the stronger PVT-v2 architecture, we achieve the best overall performance on S4 (reaching 91.63 in $\mathcal{J}\&\mathcal{F}$) and outperform ICF by 1.56 and 3.09 points on S4 and M3. Across both backbones, these consistent $\mathcal{J}$ gains validate that explicit visual candidate construction effectively suppresses irrelevant visual responses and reduces spatial ambiguity. Nevertheless, DDAVS and ICF retain marginal $\mathcal{F}$ score advantages on the more challenging M3 and AVSS benchmarks. This occurs primarily because our candidate-level abstraction

Table 5: Parameter analysis of K, Q, τ_c, τ_p, and m on S4.

Metrics	K				Q				τ_c				τ_p				m			
	4	6	8	10	1	3	5	7	0.03	0.05	0.07	0.10	0.1	0.3	0.5	0.7	0.1	0.2	0.3	0.5
$\mathcal{J}\&\mathcal{F}\uparrow$	90.55	91.44	**91.63**	90.52	90.47	90.82	**91.63**	90.63	90.51	90.76	**91.63**	90.95	90.35	90.26	**91.63**	90.47	90.73	**91.63**	90.52	90.63
$\mathcal{J}\uparrow$	88.47	89.36	**89.72**	88.51	88.35	88.55	**89.72**	88.33	88.30	88.73	**89.72**	88.95	88.35	88.21	**89.72**	88.37	88.58	**89.72**	88.44	88.50
$\mathcal{F}\uparrow$	92.63	93.52	**93.54**	92.53	92.58	93.08	**93.54**	92.93	92.71	92.78	**93.54**	92.94	92.35	92.31	**93.54**	92.56	92.88	**93.54**	92.59	92.75

inherently prioritizes robust entity localization ($\mathcal{J}$) over fine-grained boundary preservation ($\mathcal{F}$) in highly overlapping multi-source scenes. Despite this minor trade-off, our consistently strong overall performance across diverse benchmarks further demonstrates the robustness and broad applicability of the proposed asymmetric framework.

Qualitative analysis. As shown in Fig. 4, while competing methods suffer from incomplete localization or false responses in complex scenes, our approach consistently yields accurate and coherent masks. This robust performance is driven by explicit candidate construction, reliable acoustic grounding, structured response suppression, and effective counterfactual calibration.

4.3 Ablation Study and Analysis

We conduct ablation studies on S4 and M3 with PVT-v2 under identical settings. In Tables 3–4, "w/o" and "$\rightarrow$" denote removal and replacement, respectively.

Effect of the Overall Components. As shown in Table 2, progressively introducing HVCC, CF-AQR, FSC, and PCQL improves $\mathcal{J}\&\mathcal{F}$ from 86.17 to 91.63 on S4 and from 66.45 to 72.98 on M3. In particular, HVCC brings a 2.18-point gain on S4 by introducing explicit visual candidates, while CF-AQR further improves performance by 1.52 and 1.99 points on S4 and M3, respectively, demonstrating the benefit of candidate-level audio grounding. FSC provides a particularly clear improvement on the more challenging multi-source M3 setting (+1.67), indicating the effectiveness of structured semantic composition under source ambiguity. Finally, PCQL yields additional gains of 1.37 and 1.34 points on S4 and M3 through counterfactual calibration. These consistent improvements demonstrate that the four components address complementary aspects of visual entity construction, audio grounding, semantic composition, and cross-modal calibration.

Effect of Component Designs. Tables 3 and 4 detail the component-level ablations across different design choices. For HVCC, replacing learnable hierarchical weighting or soft mask pooling with uniform/hard alternatives degrades $\mathcal{J}\&\mathcal{F}$ by up to 1.01 points, validating the necessity of adaptive aggregation and soft abstraction. For CF-AQR, coarse and fine matching prove complementary; additionally, the routing gate (critical for M3, -0.85 drop) and soft verification regions consistently outperform their simplified variants. For FSC, factorised residual injection consistently surpasses direct dense fusion. Notably, removing null-mass normalization causes the largest degradation (-1.18 on S4, -0.87 on M3), highlighting its essential role in retaining implicit background mass, while stop-gradient isolation effectively prevents undesirable visual gradient interference. Finally, for PCQL, semantic-safe hard negatives outperform random ones (dropping up to 1.02 points without filtering). Progressive optimization (+1.14 on S4) and logit-space margins yield the most effective counterfactual calibration. These results validate the effectiveness of our internal structural designs.

Effect of Hyperparameters. We further analyze hyperparameter sensitivity on S4 (Table 5), where the best $\mathcal{J}\&\mathcal{F}$ (91.63) is achieved at $K = 8, Q = 5, \tau_c = 0.07, \tau_p = 0.5$, and $m = 0.2$. Insufficient candidates (K) or queries (Q) limit effective entity coverage and representation capacity, whereas excessively large values introduce unnecessary redundancy and potential optimization noise. Similarly, intermediate temperatures (τ_c, τ_p) and margin (m) achieve the best overall trade-off among discriminative matching, stable query aggregation, and robust counterfactual calibration.

5 Conclusion

In this work, we reformulate audio-visual segmentation as an asymmetric entity-querying problem, where vision constructs explicit visual candidates and audio verifies their acoustic activation. AEQNet implements this paradigm through HVCC for candidate construction, CF-AQR for coarse-to-fine audio verification, FSC for factorised semantic composition, and PCQL for suppressing visual shortcuts. Extensive experiments across multiple AVS benchmarks demonstrate the effectiveness, robustness, and generality of the proposed framework.

AI USE STATEMENT

Generative AI tools were used during manuscript revision to assist with language editing, structural organisation, and critical feedback on clarity and presentation. These tools were not treated as sources of scientific evidence and did not generate or alter the experimental data or reported results. All mathematical statements, methodological descriptions, citations, and empirical claims included in the manuscript were independently examined and verified by the authors. The authors made all final decisions and take full responsibility for the content of this work.

REFERENCES

Jack Brady, Roland S. Zimmermann, Yash Sharma, Bernhard Schölkopf, Julius Von Kügelgen, and Wieland Brendel. Provably learning object-centric representations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 3038–3062. PMLR, 2023.

Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. Unraveling instance associations: A closer look for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26497–26507, 2024a.

Yuanhong Chen, Chong Wang, Yuyuan Liu, Hu Wang, and Gustavo Carneiro. CPM: Class-conditional prompting machine for audio-visual segmentation. In *European Conference on Computer Vision*, pp. 438–456. Springer, 2024b. doi: 10.1007/978-3-031-72684-2_25.

Marco Cristani, Manuele Bicego, and Vittorio Murino. Audio-visual event recognition in surveillance video sequences. *IEEE Transactions on Multimedia*, 9(2):257–267, 2007.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In *Advances in Neural Information Processing Systems*, volume 13, 2000.

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, pp. 12155–12163, 2024. doi: 10.1609/aaai.v38i11.29104.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Yifan Wang, Pingping Zhang, Lijun Wang, and Huchuan Lu. AVS-Mamba: Exploring temporal and multi-modal mamba for audio-visual segmentation. *IEEE Transactions on Multimedia*, 27:5413–5425, 2025.

Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature Communications*, 12(1):5721, 2021.

Dawei Hao, Yuxin Mao, Bowen He, Xiaodong Han, Yuchao Dai, and Yiran Zhong. Improving audio-visual segmentation with bidirectional generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2067–2075, 2024. doi: 10.1609/aaai.v38i3.27978.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135. IEEE, 2017.

Shaofei Huang, Han Li, Yuqing Wang, Hongji Zhu, Jiao Dai, Jizhong Han, Wenge Rong, and Si Liu. Discovering sounding objects by audio queries for audio visual segmentation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 875–883, 2023. doi: 10.24963/ijcai.2023/97.

Shaofei Huang, Rui Ling, Tianrui Hui, Hongyu Li, Xu Zhou, Shifeng Zhang, Si Liu, Richang Hong, and Meng Wang. Revisiting audio-visual segmentation with vision-centric transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8352–8361, 2025.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 2021.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

Jiaxu Li, Songsong Yu, Yifan Wang, Lijun Wang, and Huchuan Lu. SelM: Selective mechanism based audio-visual segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3926–3935, 2024a.

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 *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Xiang Li, Jinglu Wang, Xiaohao Xu, Xiulian Peng, Rita Singh, Yan Lu, and Bhiksha Raj. QDFormer: Towards robust audiovisual segmentation in complex environments with quantization-based semantic decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3402–3413, 2024b.

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, 2024a. doi: 10.1109/TMM.2024.3405622.

Chen Liu, Liying Yang, Peike Li, Dadong Wang, Lincheng Li, and Xin Yu. Dynamic derivation and elimination: Audio visual segmentation with enhanced audio semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3131–3141, 2025.

Jinxiang Liu, Yikun Liu, Fei Zhang, Chen Ju, Ya Zhang, and Yanfeng Wang. Audio-visual segmentation via unlabeled frame exploitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26328–26339, 2024b.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020.

Ying Lv, Zhi Liu, and Xiaojun Chang. Consistency-queried transformer for audio-visual segmentation. *IEEE Transactions on Image Processing*, 34:2616–2627, 2025. doi: 10.1109/TIP.2025.3563076.

Ying Lv, Zhi Liu, and Xiaojun Chang. Uncertainty-aware audio-visual segmentation with dynamic fusion for multimodal alignment. *IEEE Transactions on Multimedia*, 28:3160–3173, 2026. doi: 10.1109/TMM.2026.3651123.

Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Dong Li, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *IEEE Transactions on Image Processing*, 34:4108–4119, 2025. doi: 10.1109/TIP.2025.3580269.

Kai Peng, Yunzhe Shen, Miao Zhang, Leiye Liu, Yidong Han, Wei Ji, Jingjing Li, Yongri Piao, and Huchuan Lu. Selective noise suppression and discriminative mutual interaction for robust audio-visual segmentation. *IEEE Transactions on Multimedia*, 2026. doi: 10.1109/TMM.2026.3701207.

Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1713–1721, 2015.

Nisarg A Shah, Vibashan VS, and Vishal M Patel. LQMFormer: Language-aware query mask transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12903–12913, 2024.

Li Shen, Yangzhu Wang, Xuyi Fan, Yuning Wei, and Huaxin Qiu. AV2TS: A multivariate time series modeling framework for audio-visual segmentation. *IEEE Transactions on Multimedia*, 28:4056–4068, 2026. doi: 10.1109/TMM.2026.3654457.

Jingqi Tian, Yiheng Du, Haoji Zhang, Yuji Wang, Isaac Ning Lee, Xulong Bai, Tianrui Zhu, Jingxuan Niu, and Yansong Tang. Delayed bidirectional alignment via disentangled audio semantics for audio-visual segmentation. In *European Conference on Computer Vision*, pp. 619–638. Springer, 2026.

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.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. DriveDreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pp. 55–72. Springer, 2024.

Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27134–27143, 2024.

Mingfeng Zha, Tianyu Li, Guoqing Wang, Peng Wang, Yangyang Wu, Yang Yang, and Heng Tao Shen. Implicit counterfactual learning for audio-visual segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22349–22360, 2025.

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*, pp. 386–403. Springer, 2022.

Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, 133(4):1644–1664, 2025. doi: 10.1007/s11263-024-02261-x.

Wujie Zhou, Yusen Wang, Xiaohong Qian, Yuanyuan Liu, and Lu Yu. Prompt-guided contrastive network: Efficient audio-visual segmentation via prompt learning and contrastive learning. *IEEE Transactions on Multimedia*, 2026. doi: 10.1109/TMM.2026.3714324.

Yue Zhu, Kun Li, and Zongxin Yang. Exploiting EfficientSAM and temporal coherence for audio-visual segmentation. *IEEE Transactions on Multimedia*, 27:2999–3008, 2025. doi: 10.1109/TMM.2025.3557637.

A APPENDIX

This appendix provides supplementary details and analyses for the main paper. Appendix A complements the methodology and experimental setup, including the overall procedure (Sec. 3), implementation details (Sec. 4.1), and ablation notation (Sec. 4.3). Appendix B complements the experimental results, including qualitative analysis and benchmark discussion (Sec. 4.2), detailed ablations and hyperparameter analysis (Sec. 4.3), efficiency analysis, and limitations and future work.

A.1 METHODOLOGY

The overall procedure of the proposed AEQNet is summarized in Algorithm 1. Given the visual frames and the corresponding audio signals, AEQNet first extracts hierarchical visual representations and acoustic features using separate encoders. It then constructs explicit visual candidates through HVCC, performs candidate-level acoustic verification using CF-AQR, and composes spatial membership, visual identity, and audio support through FSC for final segmentation prediction. During training, PCQL further introduces hard counterfactual audio to calibrate the audio-candidate correspondence and suppress visual shortcuts. The detailed computational procedure is presented in Algorithm 1.

A.2 IMPLEMENTATION DETAILS

For completeness, Table 6 summarizes the method-specific hyperparameters used in all experiments. Unless otherwise specified, the same configuration is used across S4, M3, and AVSS. The hyperparameters K, Q, τ_c, τ_p, and m are further examined through sensitivity analysis in Table 5. The remaining coefficients are kept fixed throughout all experiments to ensure consistent comparisons between different architectural variants. The four segmentation objectives are weighted by λ_{ce}, λ_{f}, λ_{d}, and λ_{i}, while λ_{c}, λ_{q}, and λ_{p} balance candidate construction, factual audio grounding, and counterfactual calibration, respectively. The progressive optimization starts factual query supervision at epoch $e_{\mathrm{q}} = 2$ and counterfactual calibration at epoch $e_{\mathrm{p}} = 6$, with a linear ramp of $\Delta_{\mathrm{q}} = 4$ epochs. The FSC injection coefficient follows the same schedule as PCQL, i.e., $\gamma_f(e) = \gamma_p(e)$, and therefore introduces no additional scheduling hyperparameter.

Table 6: Hyperparameter settings.

Hyperparameter	Value	Hyperparameter	Value	Hyperparameter	Value
K	8	λ_{ce}	5.0	λ_{q}	0.5
Q	5	λ_{f}	5.0	λ_{p}	0.1
τ_c	0.07	λ_{d}	5.0	e_{q}	2
τ_p	0.5	λ_{i}	2.0	e_{p}	6
m	0.2	λ_{c}	0.5	Δ_{q}	4

A.3 ABLATION NOTATION

For readability, Tables 3 and 4 use compact abbreviations for different architectural variants. Their full definitions are summarized in Table 7. Here, "w/o" denotes removing a component, whereas "$\rightarrow$" denotes replacing the default design with the alternative indicated on the right-hand side.

Algorithm 1 Asymmetric Entity-Querying Network (AEQNet)

Input: Visual frames $\mathbf{V}$ and audio $\mathbf{A}$; **Parameters:** Number of visual candidates K and audio queries Q
$\{\mathbf{F}_i^v\}_{i=1}^L \leftarrow \mathrm{VisualEncoder}(\mathbf{V}), \qquad \mathbf{F}^a \leftarrow \mathrm{AudioEncoder}(\mathbf{A})$
$\mathbf{F}^h \leftarrow \mathrm{HierarchicalAggregation}(\{\mathbf{F}_i^v\}_{i=1}^L), \qquad \mathbf{M} \leftarrow \sigma\big(G_m(\mathbf{F}^h)\big)$
for $k = 1$ to K **do**
  $\mathbf{Z}_k \leftarrow \mathrm{MaskConditionedPooling}(\mathbf{F}^h, \mathbf{M}_k)$
end for
$\mathbf{Q}^a \leftarrow \mathrm{AudioTokenEnhancement}(\mathbf{F}^a), \quad \mathbf{M}^q, \mathbf{Z}^q, \mathbf{F}^{h,q} \leftarrow sg(\mathbf{M}, \mathbf{Z}, \mathbf{F}^h)$
for $k = 1$ to K **do**
  for $q = 1$ to Q **do**
    $r_{q,k}^c \leftarrow \mathrm{CoarseSemanticMatching}(\mathbf{Q}_q^a, \mathbf{Z}_k^q)$
    $r_{q,k}^f \leftarrow \mathrm{CandidateFineVerification}(\mathbf{Q}_q^a, \mathbf{F}^{h,q}, \mathbf{M}_k^q)$
    $r_{q,k} \leftarrow \mathrm{AdaptiveRouting}(r_{q,k}^c, r_{q,k}^f)$
  end for
  $S_k \leftarrow \sigma\Big(\mathrm{LogMeanExp}(\{r_{q,k}\}_{q=1}^Q)\Big)$
end for
$\mathbf{W} \leftarrow \mathrm{NullAwareNormalization}(\mathbf{M}), \quad \mathbf{C}(u) \leftarrow \sum_{k=1}^K W_k(u)\, S_k\, \mathbf{Z}_k$
$\mathbf{F}^o \leftarrow \mathbf{F}^h + \gamma_f(e)\, Conv^{1\times1}(\mathbf{C}), \quad \hat{\mathbf{Y}} \leftarrow \mathrm{MaskDecoder}(\mathbf{F}^o)$
if training **then**
  for $b = 1$ to B **do**
    $\mathcal{N}_b \leftarrow \{j \neq b \mid \mathbf{y}_b^\top \mathbf{y}_j = 0\}, \quad j^* \leftarrow \arg\max_{j\in\mathcal{N}_b} sim(\mathbf{F}_b^a, \mathbf{F}_j^a)$
    $\ell_b^+ \leftarrow \text{CF-AQR}(\mathbf{Q}_b^a, \mathbf{M}_b^q, \mathbf{Z}_b^q), \quad \ell_b^- \leftarrow \text{CF-AQR}(\mathbf{Q}_{j^*}^a, \mathbf{M}_b^q, \mathbf{Z}_b^q)$
  end for
  $\mathcal{L}_{PCQL} \leftarrow \mathrm{MarginLoss}(\ell^+, \ell^-)$
  $\mathcal{L}_{total} \leftarrow \mathcal{L}_{seg} + \lambda_c \mathcal{L}_{cand} + \gamma_q(e)\lambda_q \mathcal{L}_{query} + \gamma_p(e)\lambda_p \mathcal{L}_{PCQL}$
  $\mathrm{Backward}(\mathcal{L}_{total})$
end if
Output: Segmentation prediction $\hat{\mathbf{Y}}$

Table 7: Notation used in the ablation studies.

Notation	Meaning	Notation	Meaning
LW	Learnable Hierarchical Weighting	CM	Coarse Semantic Matching
UA	Uniform Averaging	FV	Fine Verification
SM	Soft Mask-Conditioned Pooling	RG	Routing Gate
HM	Hard Mask Pooling	AR	Adaptive Routing
FR	Feature Refinement	LS	Linear Summation
L_{cand}	Candidate Supervision Loss	SR	Soft Verification Region
E2E	End-to-End Optimization	HR	Hard Verification Region
FC	Factorised Composition	SHN	Semantic-Safe Hard Negative
DF	Dense Fusion	SRN	Semantic-Safe Random Negative
AS	Audio Support	SSF	Semantic-Safe Filtering
RI	Residual Injection	PO	Progressive Optimization
DR	Direct Replacement	JO	Joint Optimization
NM	Null-Mass Normalization	LM	Logit-Space Margin
SN	Softmax Normalization	PM	Probability-Space Margin
SG	Stop-Gradient Isolation	L_{PCQL}	Counterfactual Query Loss

B Additional Experimental Analysis

B.1 Qualitative Component Analysis

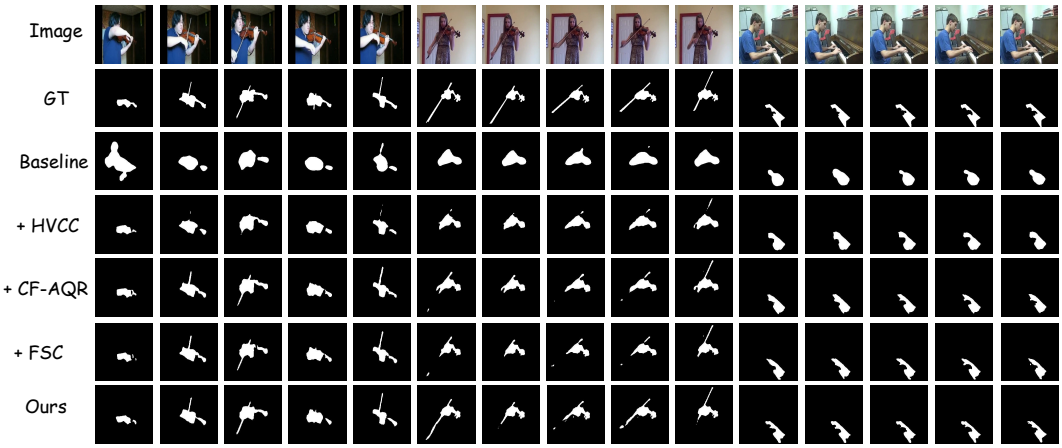

Figure 5: Qualitative evolution of segmentation predictions as the proposed components are progressively introduced. From top to bottom: input frames, ground truth, baseline, baseline + HVCC, + CF-AQR, + FSC, and the complete AEQNet with PCQL. HVCC improves object-centric localization, CF-AQR suppresses acoustically irrelevant responses, FSC produces more coherent candidate-aware predictions, and PCQL further reduces residual false activations.

Figure 5 provides a qualitative counterpart to the progressive ablation in Table 2. The baseline often captures only the most salient visual regions or produces incomplete masks. Introducing HVCC substantially improves the spatial completeness of the predicted sounding objects by explicitly constructing object-centric candidates before audio interaction. Adding CF-AQR further removes visually plausible but acoustically unsupported regions, illustrating the advantage of performing candidate-level audio verification rather than unrestricted dense cross-modal interaction. The FSC stage further improves the structural consistency of the predicted masks by combining candidate membership, visual identity, and audio support before decoding. Finally, the complete model with PCQL produces the most stable responses across frames and suppresses residual false activations. This progression is consistent with the quantitative improvements observed when the four components are introduced sequentially.

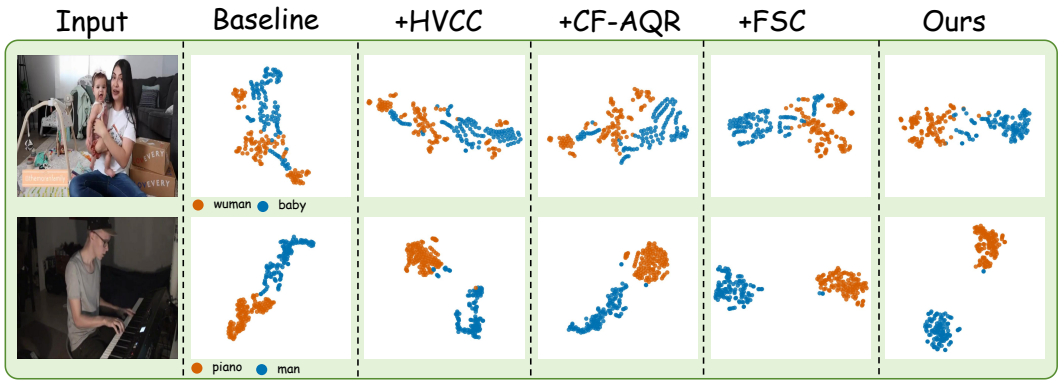

Figure 6: t-SNE visualization of feature representations as the proposed components are progressively introduced. From left to right: input, baseline, +HVCC, +CF-AQR, +FSC, and the complete AEQNet.

Figure 6 further visualizes the evolution of the learned feature space using t-SNE. Compared with the relatively entangled representations of the baseline, HVCC improves the structural organization of visual entity features, while CF-AQR further enhances their separation through candidate-level acoustic verification. FSC produces more structured feature distributions by integrating candidate membership, visual identity, and audio support. The complete AEQNet exhibits the clearest sepa-

ration and most coherent feature organization, providing a representation-level explanation for the progressive improvements observed in Table 2.

B.2 Further Discussion of Benchmark Results

The quantitative results in Table 1 reveal two consistent properties of AEQNet. First, the improvements are particularly pronounced in the Jaccard index $\mathcal{J}$, indicating that the proposed framework primarily strengthens spatial localization of sounding entities. With ResNet-50, AEQNet improves $\mathcal{J}\&\mathcal{F}$ over CPM by 3.17 and 3.66 points on S4 and M3, respectively. On M3, the $\mathcal{J}$ improvement reaches 5.12 points over CPM and 10.42 points over COMBO. These gains are consistent with the design motivation of HVCC: constructing explicit visual candidates reduces spatial ambiguity before acoustic evidence is introduced. Second, the advantage persists when switching from ResNet-50 to PVT-v2. With PVT-v2, AEQNet obtains 91.63 $\mathcal{J}\&\mathcal{F}$ on S4, together with strong performance on M3 and AVSS. This backbone consistency suggests that the proposed candidate-query-routing formulation is not tied to a particular visual encoder. We also observe that DDAVS and ICF retain slightly higher $\mathcal{F}$ scores on some challenging settings. This suggests that the candidate-level abstraction emphasizes stable region-level localization, whereas highly fine-grained boundaries in overlapping or cluttered scenes remain more challenging. Importantly, the strong $\mathcal{J}$ values across both backbones indicate that this trade-off primarily concerns boundary precision rather than the ability to identify the correct sounding entities.

B.3 Detailed Ablation Analysis

We further analyze the contribution of each component using the results in Tables 2, 3, and 4. All variants are evaluated under identical training and evaluation protocols.

B.3.1 Overall Component Contribution

Starting from the baseline, progressively adding HVCC, CF-AQR, FSC, and PCQL increases $\mathcal{J}\&\mathcal{F}$ from 86.17 to 91.63 on S4 and from 66.45 to 72.98 on M3. Importantly, every stage provides an additional improvement, indicating that the four modules address complementary sources of error rather than repeatedly modeling the same interaction.

HVCC provides the first substantial gain, improving $\mathcal{J}\&\mathcal{F}$ by 2.18 points on S4 and 1.53 points on M3. Adding CF-AQR further improves the score by 1.52 and 1.99 points, respectively. FSC contributes a further 0.39 points on S4 but a substantially larger 1.67 points on M3, suggesting that structured semantic composition is particularly useful in multi-source scenes. Finally, PCQL adds another 1.37 points on S4 and 1.34 points on M3. The consistent improvements on both datasets support the intended progression from visual entity construction to acoustic verification, structured composition, and counterfactual calibration.

B.3.2 Analysis of HVCC

Table 3 evaluates four design choices in the Hierarchical Visual Candidate Constructor.

Learnable hierarchical weighting. Replacing learnable weighting (LW) with uniform averaging (UA) reduces $\mathcal{J}\&\mathcal{F}$ from 91.63 to 91.03 on S4 and from 72.98 to 71.97 on M3. The larger 1.01-point degradation on M3 indicates that different feature levels contribute unequally when multiple sounding objects and complex spatial structures are present. Adaptive weighting therefore allows HVCC to select an appropriate balance between high-level semantics and fine spatial details.

Soft mask-conditioned pooling. Replacing soft mask pooling (SM) with hard mask pooling (HM) decreases $\mathcal{J}\&\mathcal{F}$ by 0.76 points on S4 and 0.89 points on M3. Soft masks preserve uncertainty around object boundaries and permit graded feature aggregation, whereas premature hard assignment discards potentially useful candidate evidence. This is especially important before audio verification, when the visual branch should preserve plausible entity hypotheses rather than committing to overly discrete regions.

Feature refinement. Removing feature refinement (FR) reduces performance by 0.81 points on S4 and 0.31 points on M3. This confirms that refinement of the hierarchically aggregated representa-

tion provides a cleaner basis for subsequent candidate mask prediction and candidate embedding extraction.

Candidate supervision. Removing L_{cand} decreases $\mathcal{J}\&\mathcal{F}$ by 0.45 points on S4 and 0.73 points on M3. Although the final segmentation loss can still indirectly optimize the visual branch, explicit candidate-level supervision encourages individual candidates to maintain meaningful object-centric spatial structure, which in turn provides more reliable units for subsequent audio querying.

B.3.3 Analysis of CF-AQR

Table 4 further evaluates the design of the Coarse-to-Fine Audio Query Router.

Coarse semantic matching and fine verification. Removing coarse matching (CM) decreases $\mathcal{J}\&\mathcal{F}$ by 0.83 points on S4 and 0.49 points on M3, while removing fine verification (FV) results in drops of 0.75 and 0.30 points, respectively. The two branches therefore provide complementary information: coarse matching measures global semantic compatibility between audio queries and candidate identities, whereas fine verification evaluates the consistency of audio evidence within the spatial support of each candidate.

Routing gate. Removing the routing gate (RG) reduces performance by 0.41 points on S4 and 0.85 points on M3. The substantially larger degradation on M3 suggests that adaptive balancing between coarse semantic evidence and local spatial verification becomes particularly important in multi-source scenarios, where a fixed combination can be unreliable across different candidate-audio pairs.

Adaptive routing. Replacing adaptive routing (AR) with linear summation (LS) causes a 0.70-point drop on S4. The effect on M3 is smaller (0.08 points), but the full adaptive formulation remains the strongest overall configuration. This indicates that learned pair-dependent weighting provides a more flexible integration mechanism than using a globally fixed combination.

Soft verification region. Replacing the soft verification region (SR) with a hard region (HR) reduces performance by 0.44 points on S4 and 0.18 points on M3. Similar to the HVCC pooling results, retaining soft spatial membership avoids abrupt boundary decisions and allows the audio branch to aggregate evidence from uncertain candidate regions.

B.3.4 Analysis of FSC

The FSC ablations isolate how candidate-level reasoning should be mapped back to dense visual representations.

Factorised composition. Replacing factorised composition (FC) with conventional dense fusion (DF) reduces $\mathcal{J}\&\mathcal{F}$ by 0.48 points on S4 and 0.15 points on M3. Although the numerical degradation is moderate, the result supports the central design principle that candidate membership, visual identity, and audio activation should remain explicitly distinguishable rather than being immediately re-entangled through dense multimodal fusion.

Audio support. Removing audio support (AS) decreases performance by 0.51 points on S4 and 0.18 points on M3. This verifies that candidate existence alone is insufficient: the semantic composer benefits from explicitly modulating each candidate according to its acoustic activation.

Residual injection. Replacing residual injection (RI) with direct replacement (DR) produces a larger degradation of 0.94 points on S4 and 0.63 points on M3. Preserving the original dense visual representation is therefore important, since the candidate context is intentionally compact and may omit fine spatial details. Residual injection allows candidate-level semantics to guide the dense representation without discarding its local structure.

Null-mass normalization. Replacing null-mass normalization (NM) with softmax normalization (SN) causes the largest degradation within FSC: 1.18 points on S4 and 0.87 points on M3. Standard softmax normalization forces every spatial location to be assigned across the available candidates, whereas the proposed formulation retains an implicit background mass. The result confirms that explicitly allowing "none-of-the-candidates" regions is important for suppressing irrelevant visual responses.

Stop-gradient isolation. Replacing stop-gradient isolation (SG) with fully end-to-end optimization (E2E) lowers performance by 0.88 points on S4 and 0.43 points on M3. This supports the motivation for preventing acoustic supervision from directly reshaping the visual candidate references. Gradient isolation helps preserve the intended asymmetric role assignment: visual candidates represent entity hypotheses, while the audio branch learns to verify their acoustic activation.

B.3.5 Analysis of PCQL

The final group of ablations examines how counterfactual audio samples are constructed and optimized.

Hard negative selection. Replacing semantic-safe hard negatives (SHN) with semantic-safe random negatives (SRN) reduces $\mathcal{J}\&\mathcal{F}$ by 0.63 points on S4 and 0.41 points on M3. Hard negatives provide more informative counterfactual comparisons because they are acoustically similar enough to challenge the query branch while remaining semantically incompatible with the target sample.

Semantic-safe filtering. Removing semantic-safe filtering (SSF) causes a larger drop of 1.02 points on S4 and 0.80 points on M3. This indicates that unrestricted negative sampling can introduce false negatives whose audio semantics overlap with the target foreground categories, weakening the counterfactual supervision.

Progressive optimization. Replacing progressive optimization (PO) with joint optimization (JO) reduces S4 performance by 1.14 points, while the difference on M3 is small (0.05 points). The stronger S4 degradation indicates that introducing counterfactual supervision before stable visual candidates and factual audio-candidate correspondences are formed can interfere with early optimization. The progressive schedule provides a more stable training order by first establishing visual candidates, then factual grounding, and finally counterfactual calibration.

Logit-space margin. Replacing the logit-space margin (LM) with a probability-space margin (PM) reduces performance by 0.90 points on S4 and 0.65 points on M3. Applying the margin before sigmoid avoids compression near probability extremes and provides a stronger separation signal between factual and counterfactual responses.

Counterfactual objective. Finally, removing L_{PCQL} reduces $\mathcal{J}\&\mathcal{F}$ from 91.63 to 90.26 on S4 and from 72.98 to 71.64 on M3, corresponding to drops of 1.37 and 1.34 points, respectively. The similar degradation across both datasets indicates that counterfactual calibration consistently contributes beyond the preceding candidate construction, audio routing, and semantic composition stages.

B.4 Hyperparameter Sensitivity

Table 5 studies the sensitivity to the number of visual candidates K, the number of audio queries Q, the coarse matching temperature τ_c, the pooling temperature τ_p, and the counterfactual margin m.

For K, performance increases from 90.55 to 91.63 $\mathcal{J}\&\mathcal{F}$ when the number of candidates increases from 4 to 8, but decreases to 90.52 when $K = 10$. Too few candidates limit entity coverage, whereas excessive candidates introduce redundant or fragmented hypotheses.

A similar trend is observed for Q. Increasing the number of audio queries from 1 to 5 improves $\mathcal{J}\&\mathcal{F}$ from 90.47 to 91.63, while further increasing it to 7 decreases performance to 90.63. This suggests that multiple queries are useful for representing diverse acoustic evidence, but excessive query redundancy may complicate candidate-query matching.

The temperature parameters also exhibit clear intermediate optima. $\tau_c = 0.07$ provides the strongest coarse semantic matching, while $\tau_p = 0.5$ yields the best balance when aggregating responses from multiple audio queries. Extremely low temperatures produce overly sharp matching, whereas overly high temperatures reduce discriminative contrast.

Finally, the counterfactual margin achieves its best performance at $m = 0.2$. A smaller margin provides insufficient separation between factual and counterfactual responses, while an overly large margin introduces a more aggressive constraint that can make optimization unnecessarily difficult. Overall, the performance remains competitive over a reasonable range around the selected values, while the adopted configuration provides the strongest combined $\mathcal{J}$ and $\mathcal{F}$ performance.

Table 8: Efficiency comparison with previous AVS methods on AVSS.

Metric	TPAVI	AVSegFormer	SELM	DDAVS	UAVS	AV2TS	COMBO-AVS	**Ours**
Params (M) ↓	**101.32**	113.90	114.09	150.29	161.42	197.53	201.38	104.93
FLOPs (G) ↓	**34.90**	49.00	61.71	85.72	88.84	38.42	34.94	59.39
AVSS $\mathcal{J}\&\mathcal{F}$ (%) ↑	32.50	39.33	44.10	47.55	41.60	47.20	44.10	**48.12**

B.5 Efficiency Analysis

Table 8 compares AEQNet with representative AVS methods in terms of parameter count, computational cost, and segmentation accuracy on AVSS. AEQNet achieves the highest $\mathcal{J}\&\mathcal{F}$ score of 48.12 among the compared methods while using 104.93M parameters and 59.39G FLOPs. In terms of model size, only TPAVI uses fewer parameters (101.32M), but its $\mathcal{J}\&\mathcal{F}$ performance is substantially lower. Compared with recent high-performing approaches such as DDAVS, AEQNet uses considerably fewer parameters (104.93M vs. 150.29M) while obtaining higher AVSS accuracy (48.12 vs. 47.55).

The FLOP count of AEQNet is not the lowest among all competitors, since the explicit construction and verification of visual candidates introduces additional candidate-level reasoning. Nevertheless, its computation remains below several strong alternatives, including SELM, DDAVS, and UAVS. These results indicate that the proposed asymmetric design improves segmentation performance without relying on an excessively large model, providing a favorable balance between representation capacity and model complexity.

C Limitations and Future Work

While AEQNet demonstrates strong performance across different AVS benchmarks, its candidate-level abstraction may sacrifice fine-grained boundary precision in highly overlapping or cluttered multi-source scenes. Future work will explore stronger boundary-aware candidate representations, extend asymmetric entity querying to open-vocabulary audio-visual settings, and investigate its generalization to broader multimodal localization tasks.