Can You Hear, Localize, and Segment Continually? An Exemplar-Free Continual Learning Benchmark for Audio-Visual Segmentation

Siddeshwar Raghavan[1], Gautham Vinod[1], Bruce Coburn[1], and Fengqing Zhu[1]

Purdue University, West Lafayette, IN 47906, USA
{raghav12, gvinod, coburn6, zhu0}@purdue.edu

Abstract. Audio-Visual Segmentation (AVS) aims to produce pixel-level masks of sound producing objects in videos, by jointly learning from audio and visual signals. However, real-world environments are inherently dynamic, causing audio and visual distributions to evolve over time, which challenge existing AVS systems that assume static training settings. To address this gap, we introduce the first exemplar-free continual learning benchmark for Audio-Visual Segmentation, comprising four learning protocols across single-source and multi-source AVS datasets. We further propose a strong baseline, ATLAS, which uses audio-guided pre-fusion conditioning to modulate visual feature channels via projected audio context before cross-modal attention. Finally, we mitigate catastrophic forgetting by introducing Low-Rank Anchoring (LRA), which stabilizes adapted weights based on loss sensitivity. Extensive experiments demonstrate competitive performance across diverse continual scenarios, establishing a foundation for lifelong audio-visual perception. Code is available at[*1] - https://gitlab.com/viper-purdue/atlas

Keywords: Continual Learning · Audio-Visual Segmentation · Multi-Modal Learning

1 Introduction

Humans can naturally recognize and localize objects by jointly interpreting visual cues and the sounds they produce through everyday experience [25, 31, 34]. This capability is fundamental to how we perceive, understand, and interact with complex environments. Audio-Visual Segmentation (AVS) [50, 51] targets this challenge by producing a pixel level map of the objects that produce sounds in the video frames. However, humans continuously adapt to changing surroundings, acquiring new visual and auditory patterns while retaining previously learned knowledge. Hearing a new instrument for the first time, for instance, does not

[1] Paper under review

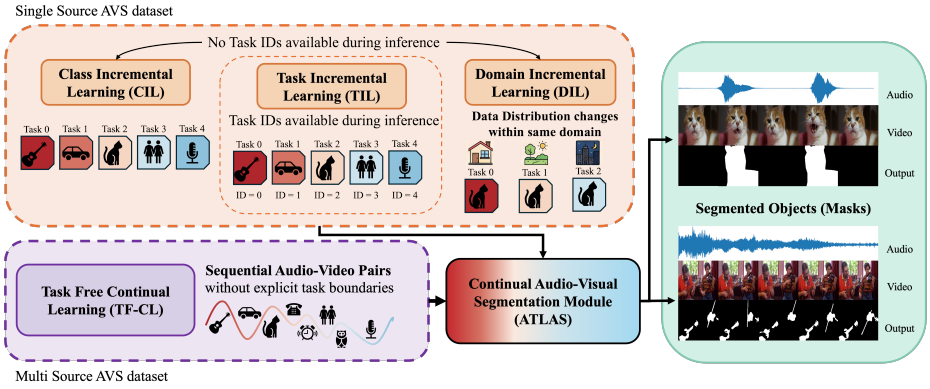

Fig. 1: An overview of Exemplar Free Continual Learning Benchmark for Audio-Visual Segmentation

cause listeners to forget the sound of a dog's bark. Replicating this lifelong learning ability in machines remains highly challenging due to the intricate correspondence and interdependence between audio and visual modalities, the scarcity of precise localization supervision, and the persistent issue of catastrophic forgetting [20, 21, 27].

Since its introduction, AVS has attracted considerable attention, with methods exploring audio-visual correspondence through transformer-based fusion [50], bidirectional cross-modal attention [15], and temporal audio-visual pixel interaction strategies [51]. Despite these advances, a critical direction remains underexplored, as real-world audio-visual perception is inherently *continual*. In practice, a deployed model encounters new sound categories over time, ranging from musical instruments and animals to vehicles and human speech. The model must learn them without retraining from scratch or revisiting prior data. Exemplar Free Continual Learning (EFCL) provides a principled framework for this setting, enabling models to acquire new knowledge sequentially while mitigating catastrophic forgetting [20, 21, 27] without storing any data. The emergence of Pre-Trained Models (PTMs) has further accelerated progress in this direction, as EFCL methods built on strong pre-trained representations [14,16,38,43,49] have demonstrated substantially better real-world applicability compared to their trained-from-scratch frameworks. While Continual Learning (CL) was originally studied in the context of image classification [1, 22, 28, 36, 37, 48], the field has since expanded considerably, including continual semantic segmentation [47], multi-modal continual learning [46], and continual reinforcement learning [35], reflecting a broader shift toward more realistic and open-ended learning scenarios.

However, directly applying these methodologies to the EFCL setting or repurposing existing methods to the EFCL setting is not trivial. AVS requires the model to simultaneously maintain cross-modal alignment between audio and vi-

sual streams, preserve fine-grained spatial segmentation boundaries, and retain previously learned sound producing object associations, all without revisiting or storing past data. The multimodal nature of AVS amplifies forgetting. Degradation in either modality or in their cross-modal alignment can cause failure, even if each modality retains useful information. Beyond forgetting, continual AVS faces several additional challenges. First, the introduction of new sounding categories causes *cross-modal interference*, where the feature spaces of both the audio and visual encoders must shift to accommodate new patterns, risking misalignment between modalities that were previously well-correlated. Second, the *evaluation complexity* increases because continual AVS requires dense pixel-level metrics that evaluate segmentation quality and audio-visual alignment while accounting for continual learning across previously encountered tasks.

To bridge this gap, we make three contributions. **First**, we introduce the exemplar-free continual learning benchmark for Audio-Visual Segmentation (CL-AVS), illustrated in Figure 1. The benchmark evaluates four continual learning protocols across two AVS datasets: Task-Incremental (TIL), Class-Incremental (CIL), and Domain-Incremental Learning (DIL) on the Single Source AVS (SS-AVS) dataset [51], and Task-Free Continual Learning [2] on the Multi-Source AVS (MS-AVS) dataset [51]. **Second**, we propose *ATLAS*, an exemplar-free baseline built on LoRA [18] adapters for parameter-efficient continual learning. It introduces an audio guided pre-fusion conditioning prior to audio–visual fusion and a Low-Rank Anchoring (LRA) mechanism to mitigate parameter drift across tasks. **Third**, we conduct extensive experiments across diverse continual learning algorithms and AVS frameworks, with detailed analysis highlighting the challenges of continual AVS and the effectiveness of our baseline

2 Related Works

2.1 Audio-Visual Segmentation

Audio-Visual Segmentation (AVS) aims to localize sound producing objects in video frames through pixel-level masks by jointly modeling audio and visual signals. The task was introduced in AVSBench [51], which established a benchmark for binary audio-visual segmentation, and later extended to semantic AVS with category-aware predictions [50]. Early approaches typically adopt encoder–decoder architectures that fuse visual representations with audio embeddings using attention or correlation mechanisms. While effective in controlled environments, these models often struggle in realistic scenes containing multiple sound sources, background noise, and visually salient but silent objects [8].

Recent AVS research broadly follows two design paradigms. **Vision-anchored** approaches leverage strong visual priors and treat audio as a conditioning signal, commonly using transformer-based fusion or vision foundation models to improve spatial precision and boundary quality [19, 26, 33]. Although effective, heavy reliance on visual cues may lead to incorrect segmentation when audio–visual correspondence weakens. In contrast, **audio-anchored** methods emphasize sound-driven localization through bidirectional interaction [15] or audio-

conditioned attention mechanisms [6], improving robustness when visual motion is ambiguous but introducing sensitivity to noisy audio. To reduce annotation cost, several works further explore weakly supervised and self-supervised learning by exploiting natural audio–visual correspondence or pseudo-label generation [29, 32].

Despite rapid progress, existing AVS methods assume static training distributions where all categories are jointly available. In real-world deployment, however, audio-visual systems encounter continuously evolving environments with new sounding categories appearing over time. To bridge this gap and set up the foundation we propose an exemplar free continual learning benchmark for AVS.

2.2 Continual Learning

Continual Learning (CL) studies learning from sequentially arriving data while mitigating catastrophic forgetting [20, 21, 27]. Early work largely focused on image classification, exploring replay-based, regularization-based, and parameter-isolation strategies [1, 22, 28, 48]. With the emergence of large pretrained models and increasing privacy constraints, recent research has shifted toward replay-free adaptation built on pretrained representations [14, 16, 43, 49], enabling scalable continual learning without storing past data.

Beyond classification, CL has expanded to dense prediction problems such as continual semantic segmentation [5, 7, 9, 45, 47] and to multimodal learning settings [38, 46]. Extending continual learning to audio-visual segmentation introduces additional challenges, as models must maintain class-level knowledge and cross-modal alignment while making fine-grained spatial predictions.

3 Methodology

In this section, we present the datasets, problem formulation for Continual Audio-Visual Segmentation, introduce the benchmark protocols defining CL-AVS, and finally describe our proposed methodology ATLAS for addressing this challenge.

3.1 Datasets

To provide the necessary context for our formal problem statement, we first describe the datasets used in this work. We utilize the datasets proposed in AVSBench [51], specifically the Single-Source AVS (SS-AVS) and Multi-Source AVS (MS-AVS) datasets. The SS-AVS consists of 4,932 videos across train, test, and validation splits with 23 categories. It is defined in a semi-supervised setting, where only the first frame of each video has ground-truth pixel level segmentation. In contrast, the MS-AVS dataset contains 424 videos covering the same 23 categories. In this setting, the videos are indexed by video ID rather than class labels, since each video clip may include multiple sound-producing objects. MS-AVS is fully supervised, with pixel-level annotations available for every frame,

but class labels are unavailable. This fundamental difference between the two datasets directly motivates the design of our continual audio-visual segmentation (CL-AVS) benchmark protocols.

3.2 Problem Formulation

The goal of audio-visual segmentation (AVS) is to learn a model capable of partitioning video frames into a set of regions that localize sound-producing objects given the corresponding audio signal. These tasks are distinguished by the details of the predicted masks. Binary AVS produces a set of foreground masks separating sounding objects from the background, while semantic AVS additionally assigns each sounding region a category label over a predefined set of object classes. Now we provide a general formulation for continual learning in AVS (CL-AVS) that unifies both settings, extending continual segmentation frameworks [4, 11] to the audio-visual domain.

CL-AVS trains the model over a sequence of N tasks $(T_1, \ldots, T_N)$, each introducing a new set of sound producing object categories. At each learning step k, a uniform number of M audio-video pairs are sampled from the dataset $\mathcal{D}$ to construct the task $T_k = \{(V_i^{(k)}, A_i^{(k)}, Y_i^{(k)})\}_{i=1}^{M}$. Here, each video is represented as a sequence of frames $V_i^{(k)} = \{v_{i,t}^{(k)}\}_{t=1}^{C_i}$ with temporally aligned audio segments $A_i^{(k)} = \{a_{i,t}^{(k)}\}_{t=1}^{C_i}$, where C_i denotes the number of frames for the i-th video. The ground-truth annotation $Y_i^{(k)} = \{y_{i,t}^{(k)}\}_{t=1}^{C_i}$ varies according to the task formulation. Let $\mathcal{K}^k$ denote the set of sounding object classes introduced at step k, and $\mathcal{K}^{1:k} = \bigcup_{j=1}^{k} \mathcal{K}^j$ the cumulative set of all classes seen up to step k, then,

$$y_{i,t}^{(k)} = \begin{cases} m_{i,t}^{(k)} \in \{0,1\}^{H \times W} & \textbf{binary setting} \\ s_{i,t}^{(k)} \in \mathcal{K}^{1:k} \cup \{0\}^{H \times W} & \textbf{semantic setting} \end{cases}$$

3.3 Continual Learning Settings

There are three major settings in Continual Learning, namely Task Incremental, Class Incremental, and Domain Incremental [40]. We adapt these settings to our benchmark CL-AVS on Single Source AVS (SS-AVS) dataset [51] and additionally extend a Task Free [23] protocol suited to Multi-Source AVS (MS-AVS) dataset [51].

Task Incremental AVS: In this setting, a new set of sound producing object classes $\mathcal{K}^k$ are introduced at each learning step k. A task identifier (ID) is provided at both training and test time, indicating which step a given sample belongs to. At test time, the model is evaluated on all seen classes $\mathcal{K}^{1:k}$ using the corresponding task identifiers.

Class Incremental AVS: Similar to the task incremental setting, new sound producing object classes $\mathcal{K}^k$ are introduced at each step k. However, no task identifier is available at test time. The model must therefore differentiate all previously seen classes $\mathcal{K}^{1:k}$ without knowledge of which step a given sample originated from.

Domain Incremental AVS: In this setting, the model operates within a single sounding object category (e.g., *barking dog*) and learns sequentially across videos of the same class but with varying data distributions, such as different visual appearances, scenes, or audio conditions. Formally, the class set remains fixed across all steps, i.e., $\mathcal{K}^1 = \mathcal{K}^2 = \cdots = \mathcal{K}^N$, while the underlying data distribution of $\mathcal{D}_k$ shifts at each step. The model is evaluated on its ability to retain segmentation performance across all encountered distributions without forgetting earlier ones.

Task-Free Continual AVS: We extend the task-free CL paradigm [23] to the multi-source AVS dataset [51], where explicit class labels are unavailable. The model observes a stream of videos across N tasks with blurry boundaries. The objective is to perform binary segmentation (sounding vs. non-sounding) without semantic distinction, evaluated over many tasks (e.g., N=50) under the anytime inference principle [23].

3.4 ATLAS Framework

We introduce Adaptive Task Learning with Anchored Stability (ATLAS), shown in Figure 2, an exemplar-free baseline for continual audio-visual segmentation. It segments sound-producing objects from video frames across sequential tasks without retraining or revisiting past data. Following the problem formulation, at a given task $T_k = \{(V_i^{(k)}, A_i^{(k)}, Y_i^{(k)})\}_{i=1}^{M}$, ATLAS maps each pair $(V_i^{(k)}, A_i^{(k)})$ to mask logits and class logits (optionally) while constraining parameter drift across tasks to minimize catastrophic forgetting during continual learning. Let the visual encoders and audio encoders be $\Phi(\cdot;\Theta_v)$ and $\Phi(\cdot;\Theta_a)$, respectively. To enable parameter efficient adaptation, we train selected linear maps in the visual encoder backbone with LoRA [18] adapters. For each pretrained matrix W_0, the updated weight is defined as:

$$W = W_0 + \Delta W \text{ and } \Delta W = \frac{\alpha}{r} BA \tag{1}$$

where $A \in \mathbb{R}^{r \times d_{in}}$ and $B \in \mathbb{R}^{d_{out} \times r}$ with rank $r \ll min(d_{in}, d_{out})$. The hyperparameter r acts as a constant scaling factor. The LoRA adapter effectively update the weights of the visual encoder to produce adapted visual features. For each sample i, the visual and audio encoded features are

$$X_{v,i}^{(k)} = \Phi_v(V_i^{(k)}; W) \text{ and } X_{a,i}^{(k)} = \Phi_a(A_i^{(k)}; \Theta_a) \tag{2}$$

These visual encoded features then flow into the **Audio Guided Pre-Fusion Conditioning** module which injects global audio context into visual tokens

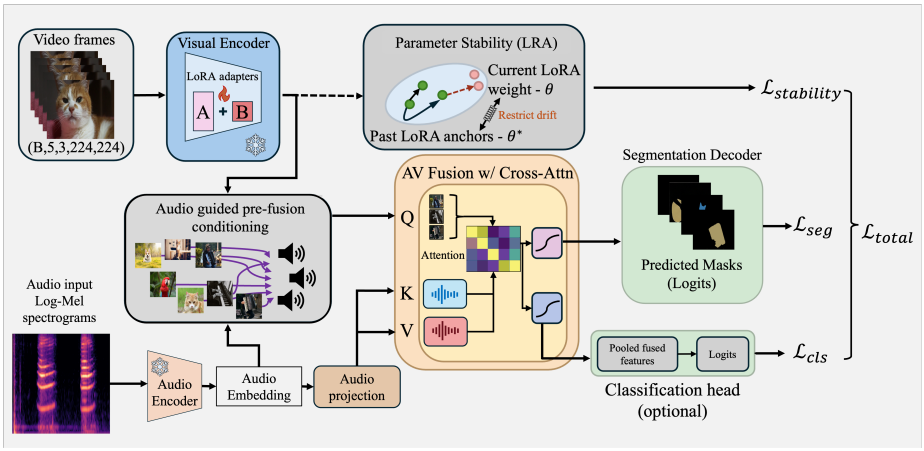

Fig. 2: Overview of ATLAS: The framework performs exemplar-free continual audio-visual segmentation using frozen encoder backbones with parameter-efficient LoRA adapters in the visual pathway. Audio-guided pre-fusion conditioning is followed by cross-attention, where conditioned visual features act as queries and audio features serve as keys and values. To mitigate catastrophic forgetting across tasks, our Low-Rank Anchoring (LRA) module dynamically restricts LoRA parameter drift. The fused representation is decoded for segmentation with an optional classification branch.

via learned channel wise modulation $\bar{X}_{v,i}^{(k)} = \mathcal{C}(X_{v,i}^{(k)}, X_{a,i}^{(k)})$. The audio-guided conditioning operator is $\mathcal{C}(\cdot)$ (detailed steps provided in supplementary material). Intuitively, this learned channel-wise modulation acts as a feature-level gating mechanism. Here, the audio features are projected to the visual token space to produce modulation parameters and conditioned tokens $\bar{X}_{v,i}^{(k)}$. By projecting the audio context into scaling and shifting parameters, the network selectively amplifies the specific visual channels that correspond to the sound-producing object while suppressing irrelevant background noise. Before AV fusion, this preliminary step aligns the visual features with sound relevant regions. These pre-conditioned visual features in turn serve as the queries (Q) in the subsequent **AV Fusion with Cross-Attention** module. We compute the Q, K, and V from these pre-conditioned visual features and the audio features respectively:

$$Q_i^{(k)} = W_Q \bar{X}_{v,i}^{(k)}, \quad K_i^{(k)} = W_K X_{a,i}^{(k)}, \quad V_i^{(k)} = W_V X_{a,i}^{(k)}. \tag{3}$$

To preserve the spatial integrity of the visual features during fusion, the fused features are obtained by taking softmax over the scaled dot product attention and applying a residual connection with Layer Normalization [41]:

$$F_i^{(k)} = \texttt{LayerNorm}\left(\bar{X}_{v,i}^{(k)} + \texttt{softmax}\left(\frac{Q_i^{(k)} K_i^{(k)^T}}{\sqrt{d}}\right) V_i^{(k)}\right) \tag{4}$$

The decoder $\mathcal{D}$, which also contains the LoRA adapted weights to mitigate catastrophic forgetting, projects these fused multimodal features to the resolution of

the original video frames to output the mask logits:

$$\hat{M}_i^{(k)} = \mathcal{D}\left(F_i^{(k)}\right) \tag{5}$$

Additionally, to support the prediction of class logits, a global pooling layer followed by a linear classification head $\mathcal{H}$ processes the combined representation to output the class logits

$$\hat{z}_i^{(k)} = \mathcal{H}\left(F_i^{(k)}\right) \tag{6}$$

For the binary setting in our CL-AVS benchmark, $\left(y_{i,t}^{(k)} = m_{i,t}^{(k)}\right)$, segmentation is optimized using a combination of Binary Cross Entropy and Dice [39] losses to handle severe foreground to background imbalances, in a supervised manner:

$$\mathcal{L}_{seg}^{(k)} = \frac{1}{M}\sum_{i=1}^{M}\left[\mathcal{L}_{BCE}\left(\hat{M}_i^{(k)}, m_i^{(k)}\right) + \mathcal{L}_{Dice}\left(\hat{M}_i^{(k)}, m_i^{(k)}\right)\right] \tag{7}$$

And for the semantic setting where we predict the logits in a supervised way over the classes seen so far $Q^{1:k}$, the classification term is standard Cross Entropy (CE):

$$\mathcal{L}_{cls}^{(k)} = \frac{1}{M}\sum_{i=1}^{M}\texttt{CE}\left(\hat{z}_i^k, c_i^{(k)}\right) \tag{8}$$

where $c_i^{(k)} \in Q^{1:k}$ is the sample level class label. Finally, we introduce the **Low-Rank Anchoring (LRA)** module to minimize drift between the past anchor weights and current weights to mitigate catastrophic forgetting. Let θ denote the current trainable parameters, θ^* the previous task's anchor weights, and Ω_i the parameter importance weight. Rather than computing static Fisher approximations, Ω_i is computed dynamically during training by accumulating the product of parameter gradients and their updates. This effectively tracks loss sensitivity along the optimization trajectory. The stability regularization is applied to LoRA matrices and decoder weights. The stability term is defined as:

$$\mathcal{L}_{stab} = \frac{c}{2}\sum_i \Omega_i\left(\theta_i - \theta_i^*\right)^2 \tag{9}$$

where c is a hyperparameter controlling the regularization strength. Our final objective at any task k is the combination of losses defined in equations 7,8,9

$$\mathcal{L}_{total}^{(k)} = \mathcal{L}_{seg}^{(k)} + \lambda_{cls}\mathcal{L}_{cls}^{(k)} + \mathcal{L}_{stab}^{(k)} \tag{10}$$

4 Experimentation Setup

4.1 Implementation Details

To ensure rigorous evaluation, we select representative AVS and continual learning methods, prioritizing publicly available code and strong reported perfor-

mance. For non-public implementations (marked with $*$), we reconstructed models using LLM-assisted development (Anthropic Claude Code [3] and Google Gemini [12]).

Our experimental setup spans four closely related categories of methods adaptable to end-to-end CL-AVS training: (1) (■) Exemplar-free continual learning methods adapted to AVS by replacing (or adding) classification heads with segmentation heads along with audio encoder; (2) (■) Static AVS models trained with continual data streams; (3) (■) continual semantic segmentation frameworks extended with audio encoders and MSE-based cross-modal alignment; and (4) (■) the replay-based CL-AVS baseline [17].

Across our experiments we use use PANN [24] (trained on AudioSet [13]) as the audio encoder. Visual backbones follow their original designs: most methods, including ATLAS, use a pretrained ViT-B/16 [10], while AVSBench, AVSBidirectional, and AVSegFormer, AVS-VCT [19] retain PVT-v2 backbones [42]. For ATLAS, we set $r = 8$, $\alpha = 16$, $\lambda_{cls} = 0.5$, $c = 0.3$. We train all methods for 30 epochs using AdamW with a learning rate 10^{-3} and weight decay 5×10^{-4}. SS-AVS follows the 11-2 split (7 tasks total), and MS-AVS uses the 31-5 split (50 tasks). Results are averaged over five runs with different seeds on four NVIDIA A40 GPUs.

4.2 Evaluation Protocol and Metrics

We follow the standard continual learning evaluation protocol and continual semantic segmentation metrics [30, 40, 47]. The validation set is used exclusively for model selection, while the held out test set is used to calculate the accuracy matrix $\mathbf{A}$ from which all continual learning metrics are measures. After training on task t, the model is evaluated on all previously seen tasks ($k \leq t$) and one unseen task (when available) to measure forward transfer. Task identity is provided during evaluation in Task-Incremental Learning (TIL), while Class-Incremental, Domain-Incremental and Task-Free settings require inference without task information.

1. **Segmentation Metrics:**
   (a) **Mean Average Precision (mAP)** Measures the ranking quality of confidence scores independent of a fixed threshold (↑ better).

$$\mathrm{AP} = \sum_j \mathrm{Prec}(j), \Delta\mathrm{Rec}(j), \qquad \mathrm{mAP} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} \mathrm{AP}_i. \tag{11}$$

   (b) **Maximum F-score (Max-F)** Captures the best achievable binary segmentation performance across decision thresholds (↑ better).

$$F_1(\tau) = \frac{2P(\tau)R(\tau)}{P(\tau) + R(\tau)}, \qquad \text{Max-F} = \max_{\tau \in [0,1]} F_1(\tau). \tag{12}$$

2. **Continual Learning Metrics:**

Accuracy Matrix A Let T denote the total number of incremental tasks. After learning task t, evaluation on task k yields

$$\mathbf{A} \in \mathbb{R}^{T\times T}, \qquad A_{t,k} = \text{performance after task } t \text{ evaluated on task } k. \quad (13)$$

This matrix summarizes performance across the task sequence.

(a) **Learning Accuracy (LA)** Performance when each task is first learned (↑ better).

$$\text{LA} = \frac{1}{T}\sum_{t=0}^{T-1} A_{t,t}. \quad (14)$$

(b) **Average Accuracy (AA)** Final performance over all tasks after completing training (↑ better).

$$\text{AA} = \frac{1}{T}\sum_{k=0}^{T-1} A_{T-1,k}. \quad (15)$$

(c) **Forgetting (F)** Measures degradation on earlier tasks after learning new ones (↓ better).

$$\text{F} = \frac{1}{T-1}\sum_{k=0}^{T-2}\left(\max_{t\in k,\ldots,T-1} A_{t,k} - A_{T-1,k}\right). \quad (16)$$

(d) **Backward Transfer (BWT)** Influence of learning new tasks on previously learned tasks (positive ↑ desirable).

$$\text{BWT} = \frac{1}{T-1}\sum_{k=0}^{T-2}\left(A_{T-1,k} - A_{k,k}\right). \quad (17)$$

(e) **Forward Transfer (FWT)** Zero-shot generalization to future tasks before training (positive ↑ desirable).

$$\text{FWT} = \frac{1}{T-1}\sum_{k=1}^{T-1} A_{k-1,k}. \quad (18)$$

Additional segmentation metrics (AUPR, mIoU, and Dice) are reported in the supplementary material.

5 Results and Discussion

Table 1 presents results across the SS-AVS benchmark (TIL, CIL, DIL; 11-2 split, 7 tasks) and MS-AVS benchmark (TF-CL; 31-5 split, 50 tasks). ATLAS achieves the highest mAP in all four settings, outperforming the runner-up by 7 to 17 points. Among exemplar-free CL methods (■), adapted to AVS by adding segmentation decoders and a PANN [24] audio encoder, the regularization-based

		SS-AVS (11-2 split)						MS-AVS (31-5 split)	
		TIL		CIL		DIL		TF-CL	
		Avg mAP	Avg For	Avg mAP	Avg For	Avg mAP	Avg For	Avg mAP	Avg For
	Finetune	27.91	7.85	28.04	10.89	25.38	9.30	16.61	11.58
	LwF [28] [TPAMI '16]	26.70	10.34	26.89	7.52	24.79	12.52	21.59	13.43
	EWC [22] [PNAS '17]	54.90	23.12	55.03	23.52	41.37	23.47	18.89	31.81
	SI [48] [ICML '17]	59.10	13.74	56.52	15.57	42.44	22.77	19.51	33.25
	MAS [1] [ECCV '18]	56.92	20.86	56.39	20.23	42.97	25.37	26.25	26.09
	L2P [43] [CVPR '21]	28.80	10.59	25.21	12.97	23.15	9.73	20.76	15.87
	RanPAC [49] [NIPS '23]	29.36	0.00	27.80	0.00	43.99	0.00	27.62	0.00
	FeCAM [14] [NIPS '23]	28.49	0.00	28.07	0.00	44.17	0.00	27.33	0.00
	DGR [16] [CVPR '24]	31.32	40.59	28.77	34.96	45.92	13.09	36.95	19.61
	PANDA [38] [AAAI '26]	33.13	35.34	31.15	27.85	46.12	12.44	38.17	18.03
	AVSBench [50] [ECCV '22]	63.84	22.05	62.28	23.64	52.78	24.95	31.15	21.09
	AVS-Bidirection [15] [AAAI '24]	50.27	26.95	50.49	26.19	43.20	24.12	31.69	26.32
	COMBO [44] [CVPR '24]	27.67	7.38	27.18	8.89	23.07	9.95	23.76	11.02
	AVS-VCT [19] [CVPR '25]	25.64	8.60	25.54	11.18	22.19	13.49	21.98	13.87
	BalConpas [5] [ECCV '24]	23.71	10.84	22.90	10.07	21.68	11.44	20.16	16.92
	EIR [45] [CVPR '25]	29.54	9.12	28.91	10.44	26.70	9.73	22.77	10.71
	CSTA* [7] [TCSV '25]	20.68	1.31	20.41	0.24	19.32	0.16	14.50	0.60
	HVPL* [9] [ICCV '25]	28.31	0.98	32.68	19.04	35.97	1.56	26.50	0.26
	CMR* [17] [arXiV '25]	22.75	0.32	19.16	2.77	19.59	1.30	25.92	1.35
	ATLAS*(Ours)* (*Ours*)	**74.67**	9.42	**71.39**	10.14	**63.84**	13.96	**45.27**	23.71

Table 1: Performance on the Continual AVS Benchmark exemplar-free CL methods adapted to AVS; static AVS models extended to continual learning; continual segmentation adapted to AVS; the CL-AVS baseline with replay; and our method. Performance is measured by mAP across tasks and average backward forgetting.

EWC [22], SI [48], and MAS [1] perform best on SS-AVS by penalizing changes to task-important weights via Fisher information, trajectory-based accumulation, and output sensitivity, respectively. However, their per-weight scalar importance degrades on MS-AVS with 50 tasks, where accumulated penalties conflict with large number of tasks and multiple sounding objects. L2P [43] steers a frozen ViT via a prompt pool, but its prompts operate only on visual tokens, leaving the audio branch with no pathway to influence the prompted representations. RanPAC [49] and FeCAM [14] have zero forgetting by freezing features entirely, but their mAP remains near the Finetune baseline. This confirms that freezing visual features without adaptive cross-modal alignment cannot help localize sounding objects. CMR [17] ( ) replays stored exemplars but achieves only 22.75 (TIL) mAP, indicating that naive replay without explicit audio-visual grounding is insufficient.

Among static AVS models ( ) trained continually, AVSBench [50] performs best on PVT-v2 due to its dedicated audio-visual interaction module, but lacks a CL mechanism and suffers from high forgetting as continuous fine-tuning overwrites learned fusion weights. AVS-Bidirection [15] performs worse due to bidirectional attention, as parameter updates go through both visual and audio directions and increase forgetting. COMBO [44] and AVS-VCT [19] hover near the Finetune baseline, suggesting their single-task design improvements lead to no continual learning improvements. Continual segmentation methods ( ) adapted with audio encoders and MSE for alignment exhibit a pattern. HVPL [9] and

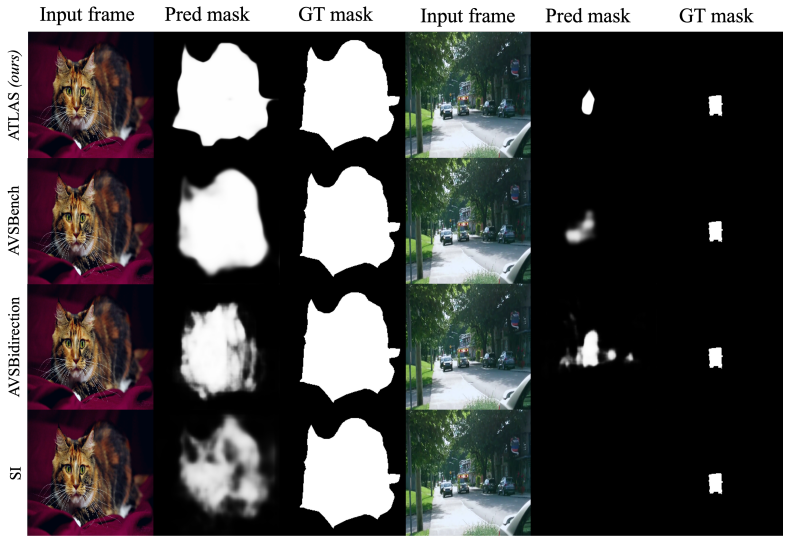

Fig. 3: Qualitative AVS results from the top four methods in our CL-AVS setting: Input frame, predicted binary mask for the middle frame is shown along with the ground truth mask. **(Left)** SS-AVS CIL and **(Right)** MS-AVS TF-CL

CSTA [7] achieve very low forgetting but fair mAP, as their single modality design updates too few parameters to learn the cross-modal correspondences AVS needs. DGR [16] and PANDA [38] perform well through gradient reweighting and patch-distribution mechanisms. Their high forgetting rates reveal their shortcomings in adapting to two modalities. We attribute the gains of our proposed model, ATLAS, to our major components. Firstly, LoRA adapters keep updates within the low-rank subspaces of the visual encoder and decoder. Followed by audio-guided pre-fusion conditioning that leads visual features toward sound-relevant regions before performing cross-attention. Finally, our Low-Rank Anchoring dynamically penalizes adapter and decoder weight drift from previous-task anchors. Additionally, Figure 3 shows qualitative segmentation results for the four best-performing methods, highlighting the strength of our approach.

Figure 4 plots Forward Transfer vs. Forgetting for TIL and DIL in the SS-AVS dataset with 11-2 split. ATLAS achieves the highest forward transfer with moderate forgetting. Most methods cluster at low FWT (< 0.45), either forgetting heavily (DGR, PANDA, MAS) or preserving knowledge at the cost of plasticity (CSTA, FeCAM, RanPAC). Under DIL, the gap is significant. ATLAS exceeds 0.55 FWT while maintaining comparable forgetting rates to those of lower-performing models.

Figure 5 shows MaxF heatmaps for the CIL setting in SS-AVS. ATLAS maintains bright cells across tasks, with early tasks remaining strong even by row 6, suggesting that our Low-Rank Anchoring mechanism effectively limits drift from previously learned representations. AVSBench performs well on early tasks but

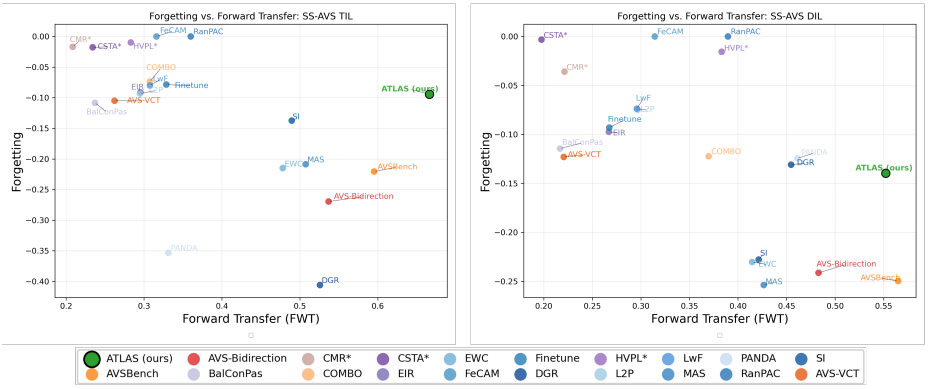

Fig. 4: Forward Transfer vs. Forgetting: Trade-off between forward transfer and forgetting across models in the CL-AVS benchmark on the SS-AVS dataset under the TIL and DIL protocols.

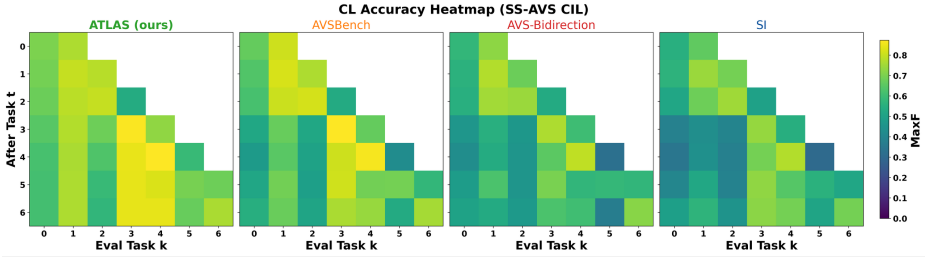

Fig. 5: Heatmap of MaxF scores on the SS-AVS test set under the CIL protocol with an 11–2 split, evaluated after each incremental task. Four best-performing methods are included for clarity.

gradually reduces as new tasks are introduced since it is not designed for continual learning. AVS-Bidirectional shows a sharp drop after task 3, likely due to error propagating through its bidirectional interactions. In contrast, SI preserves individual parameter importance but struggles to maintain the distributed cross-modal representations required for accurate segmentation. Further results and visualizations are provided in the supplementary material, including hyperparameter analyzes, additional metrics, and experiments with alternative backbone architectures. We note that the performance of the continual learner depends strongly on the quality of the pretrained models.

6 Ablation Studies

In this section, we conduct ablation studies analyzing efficiency–performance trade-offs and the contributions of individual components of our method. Figure 6 shows normalized radar profiles across six metrics: Accuracy, Retention,

			SS-AVS (11-2 split)				MS-AVS (31-5 split)			
Method	Pre-cond.	LRA	Avg mAP	Avg For	Avg F1	Avg mIoU	Avg mAP	Avg For	Avg F1	Avg mIoU
ATLAS	✓	✓	**74.67**	**9.42**	**62.76**	**50.85**	**47.53**	23.71	**42.27**	**39.44**
ATLAS	✗	✓	71.17	12.11	59.97	47.78	43.12	25.07	39.38	38.2 7
ATLAS	✓	✗	67.18	18.03	50.38	39.16	34.41	26.61	36.72	24.12
ATLAS	✗	✗	65.33	21.00	47.31	33.58	29.19	27.01	31.16	20.12
AVSBench	-	-	63.84	22.05	50.63	37.28	31.15	**21.09**	34.39	20.15

Table 2: Component ablation of ATLAS. We analyze the contribution of audio-guided pre-conditioning and Low-Rank Anchoring (LRA) on the CL-AVS benchmark

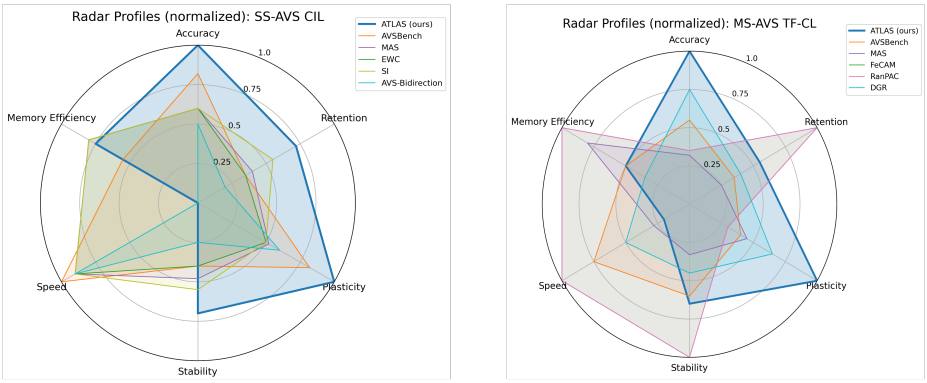

Fig. 6: Radar plot summarizing normalized performance across six dimensions: Accuracy, Retention, Plasticity, Stability, Speed, and Memory Efficiency for the top six performing methods on the SS-AVS dataset under the CIL protocol and the MS-AVS dataset under the TF-CL protocol.

Plasticity, Stability, Speed, and Memory Efficiency. ATLAS covers the largest area for both SS-AVS (CIL) and MS-AVS (TF-CL), showing strong performance particularly in Accuracy and Plasticity. Regularization methods exhibit smaller overall coverage, while prototype based approaches such as RanPAC and FeCAM are memory efficient but tend to underfit. To understand the contribution of each component of ATLAS, we perform ablations summarized in Table 2, using AVSBench as a baseline for comparison. Results show that Low-Rank Anchoring (LRA) is the most critical component, as it stabilizes LoRA-adapted weights and reduces parameter drift during continual learning. Audio-guided pre-fusion conditioning provides additional gains but contributes less than LRA. Even without these modules, the LoRA-based model suffers from parameter drift across tasks, resulting in performance below AVSBench primarily on MS-AVS larger number of tasks.

7 Conclusion

This paper introduces an exemplar-free continual learning benchmark for audio-visual segmentation, aimed at producing pixel-level masks of sounding objects

in videos. We evaluate a diverse set of methods across four continual learning protocols on both Single-Source and Multi-Source AVS datasets. We further propose ATLAS, an end-to-end exemplar-free baseline that improves learning while reducing representation drift. Our experiments demonstrate its effectiveness and establish a foundation for future research in continual audio-visual learning.

References

1. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., Tuytelaars, T.: Memory aware synapses: Learning what (not) to forget (2018), https://arxiv.org/abs/1711.09601
2. Aljundi, R., Kelchtermans, K., Tuytelaars, T.: Task-free continual learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
3. Anthropic: Introducing the next generation of claude. https://www.anthropic.com/news/claude-3-family (2024)
4. Cermelli, F., Cord, M., Douillard, A.: Comformer: Continual learning in semantic and panoptic segmentation. In: IEEE/CVF Computer Vision and Pattern Recognition Conference (2023)
5. Chen, J., Cong, R., Luo, Y., Ip, H.H.S., Kwong, S.: Strike a balance in continual panoptic segmentation. In: ECCV (2024)
6. Chen, T., Tan, Z., Chu, Q., Wu, Y., Liu, B., Yu, N.: Bootstrapping audio-visual segmentation by strengthening audio cues. In: International Conference on Learning Representations (ICLR) Workshop (2024), https://openreview.net/forum?id=WdWGe88RdX, addresses modality imbalance by enabling deeper bidirectional interaction between audio and visual features — an *audio-anchored interaction* perspective.
7. Chen, T., Liu, H., Lim, C.H., See, J., Gao, X., Hou, J., Lin, W.: Csta: Spatial-temporal causal adaptive learning for exemplar-free video class-incremental learning. IEEE Transactions on Circuits and Systems for Video technology **35**(11), 11488–11501 (2025)
8. Chen, Y., Liu, Y., Wang, H., Liu, F., Wang, C., Frazer, H., Carneiro, G.: Unraveling instance associations: A closer look for audio-visual segmentation. arXiv preprint arXiv:2304.02970 (2023), https://arxiv.org/abs/2304.02970, analysis of cross-modal alignment and benchmark bias in AVS
9. Dong, J., Yin, H., Liang, W., Zhao, H., Ding, H., Sebe, N., Khan, S., Khan, F.S.: Hierarchical visual prompt learning for continual video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2025)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
11. Douillard, A., Chen, Y., Dapogny, A., Cord, M.: Plop: Learning without forgetting for continual semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
12. Gemini Team, Google DeepMind: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
13. Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780 (2017). https://doi.org/10.1109/ICASSP.2017.7952261
14. Goswami, D., Liu, Y., Twardowski, B., van de Weijer, J.: Fecam: exploiting the heterogeneity of class distributions in exemplar-free continual learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)

15. Hao, D., Mao, Y., He, B., Han, X., Dai, Y., Zhong, Y.: Improving audio-visual segmentation with bidirectional generation. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023), https://arxiv.org/abs/2308.08288
16. He, J.: Gradient reweighting: Towards imbalanced class-incremental learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 16668–16677 (June 2024)
17. Hong, Y., Yang, Q., Zhang, T., Wang, Z., Fu, Z., Ding, K., Fan, B., Xiang, S.: Taming modality entanglement in continual audio-visual segmentation (2025), https://arxiv.org/abs/2510.17234
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), https://arxiv.org/abs/2106.09685
19. Huang, S., Ling, R., Hui, T., Li, H., Zhou, X., Zhang, S., Liu, S., Hong, R., Wang, M.: Revisiting audio-visual segmentation with vision-centric transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025), https://openaccess.thecvf.com/content/CVPR2025/papers/Huang_Revisiting_Audio-Visual_Segmentation_with_Vision-Centric_Transformer_CVPR_2025_paper.pdf, proposes a *vision-centric* transformer where visual queries guide audio fusion, improving contour and separation accuracy.
20. Kemker, R., McClure, M., Abitino, A., Hayes, T., Kanan, C.: Measuring catastrophic forgetting in neural networks. Proc. Conf. AAAI Artif. Intell. **32**(1) (Apr 2018)
21. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. U. S. A. **114**(13), 3521–3526 (Mar 2017)
22. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017). https://doi.org/10.1073/pnas.1611835114, https://www.pnas.org/doi/abs/10.1073/pnas.1611835114
23. Koh, H., Kim, D., Ha, J.W., Choi, J.: Online continual learning on class incremental blurry task configuration with anytime inference (2022), https://arxiv.org/abs/2110.10031
24. Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D.: Panns: Large-scale pretrained audio neural networks for audio pattern recognition (2020), https://arxiv.org/abs/1912.10211
25. Leavitt, V.M., Molholm, S., Gomez-Ramirez, M., Foxe, J.J.: "what" and "where" in auditory sensory processing: a high-density electrical mapping study of distinct neural processes underlying sound object recognition and sound localization. Front. Integr. Neurosci. **5**, 23 (Jun 2011)
26. Li, J., Zhao, W., Huang, Z., Guo, Y., Tian, Y.: Do audio-visual segmentation models truly segment sounding objects? arXiv preprint arXiv:2502.00358 (2025), https://arxiv.org/abs/2502.00358, investigates visual dominance bias and robustness under negative audio.
27. Li, X., Zhou, Y., Wu, T., Socher, R., Xiong, C.: Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 3925–3934. PMLR (09–15 Jun 2019), https://proceedings.mlr.press/v97/li19m.html

28. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**, 2935–2947 (2016), https://api.semanticscholar.org/CorpusID:4853851
29. Liu, J., et al.: Annotation-free audio-visual segmentation. WACV (2024), https://openaccess.thecvf.com/content/WACV2024/papers/Liu_Annotation-Free_Audio-Visual_Segmentation_WACV_2024_paper.pdf, synthesizes AVS datasets via matching image masks and audio with minimal annotation and adapts SAM to AVS.
30. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning (2022), https://arxiv.org/abs/1706.08840
31. Mihalik, A., Noppeney, U.: Causal inference in audiovisual perception. Journal of Neuroscience **40**(34), 6600–6612 (2020). https://doi.org/10.1523/JNEUROSCI.0051-20.2020, https://www.jneurosci.org/content/40/34/6600
32. Mo, S., Raj, B.: Weakly-supervised audio-visual segmentation. In: NeurIPS Workshop on Audio-Visual Machine Learning (2023), https://openreview.net/forum?id=sUqG96QqZM, proposes weakly-supervised AVS with multi-scale contrastive learning
33. Mo, S., Tian, Y.: Av-sam: Segment anything model meets audio-visual localization and segmentation. arXiv preprint arXiv:2305.01836 (2023), https://arxiv.org/abs/2305.01836, explores integration of SAM with audio-visual segmentation via pixel-wise fusion, representing an early *vision-anchored foundation model* approach.
34. Newell, F.N., McKenna, E., Seveso, M.A., Devine, I., Alahmad, F., Hirst, R.J., O'Dowd, A.: Multisensory perception constrains the formation of object categories: a review of evidence from sensory-driven and predictive processes on categorical decisions. Philos. Trans. R. Soc. Lond. B Biol. Sci. **378**(1886), 20220342 (Sep 2023)
35. Pan, C., Yang, X., Li, Y., Wei, W., Li, T., An, B., Liang, J.: A survey of continual reinforcement learning (2025), https://arxiv.org/abs/2506.21872
36. Raghavan, S., He, J., Zhu, F.: Delta: Decoupling long-tailed online continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
37. Raghavan, S., He, J., Zhu, F.: Online class-incremental learning for real-world food image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 8195–8204 (January 2024)
38. Raghavan, S., He, J., Zhu, F.: Panda – patch and distribution-aware augmentation for long-tailed exemplar-free continual learning. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025), https://arxiv.org/abs/2511.09791
39. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support (2017) **2017**, 240–248 (Sep 2017)
40. Sun, H.L., Zhou, D.W., Ye, H.J., Zhan, D.C.: Pilot: A pre-trained model-based continual learning toolbox. arXiv preprint arXiv:2309.07117 (2023)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
42. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media **8**(3), 415–424 (Sep 2022). https://doi.org/10.1007/s41095-022-0274-8, http://dx.doi.org/10.1007/s41095-022-0274-8

43. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 139–149 (2022)
44. Yang, Q., Nie, X., Li, T., Gao, P., Guo, Y., Zhen, C., Yan, P., Xiang, S.: Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation (2024)
45. Yin, H., Feng, T., Lyu, F., Shang, F., Liu, H., Feng, W., Wan, L.: Beyond background shift: Rethinking instance replay in continual semantic segmentation (2025), https://arxiv.org/abs/2503.22136
46. Yu, D., Zhang, X., Chen, Y., Liu, A., Zhang, Y., Yu, P.S., King, I.: Recent advances of multimodal continual learning: A comprehensive survey (2024), https://arxiv.org/abs/2410.05352
47. Yuan, B., Zhao, D.: A survey on continual semantic segmentation: Theory, challenge, method and application. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10891–10910 (Dec 2024). https://doi.org/10.1109/tpami.2024.3446949, http://dx.doi.org/10.1109/TPAMI.2024.3446949
48. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 3987–3995. ICML'17, JMLR.org (2017)
49. Zhou, D.W., Ye, H.J., Zhan, D.C., Liu, Z.: Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. Proceedings of the Neural Information Processing Systems (2023)
50. Zhou, J., Shen, X., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-visual segmentation with semantics. International Journal of Computer Vision pp. 1–21 (2024)
51. Zhou, J., Wang, J., Zhang, J., Sun, W., Zhang, J., Birchfield, S., Guo, D., Kong, L., Wang, M., Zhong, Y.: Audio-visual segmentation. In: European Conference on Computer Vision (2022)