Listen, Look, and Learn: Learning Without Forgetting through SAM-Audio

Avi Gupta [1] **Nilotpal Sinha** [2] **Vishnu Raj** [2] **Sambuddha Saha** [2] **Pratik Joshi** [2] **Koteswar Rao Jerripothula** [3 1]
Tammam Tillo [4 1]

Abstract

Class-Incremental Learning (CIL) aims to continuously learn new classes without forgetting previously acquired knowledge. While recent CIL advances have spurred significant interest across various modalities, the audio-visual setting remains underexplored. Furthermore, although foundational multimodal models like SAM-Audio encapsulate rich static priors, our empirical analysis reveals that these representations struggle in incremental settings. This work bridges this gap by integrating SAM-Audio's audio-visual priors into the CIL setting. Specifically, we leverage its dense audio and visual representations and employ a novel guided attention strategy where the audio features contextually guide the visual representations. To further mitigate catastrophic forgetting, we introduce dual-level distillation objectives at both the feature and logit levels. Extensive evaluations on audio-visual CIL benchmarks demonstrate that our approach consistently outperforms state-of-the-art methods.

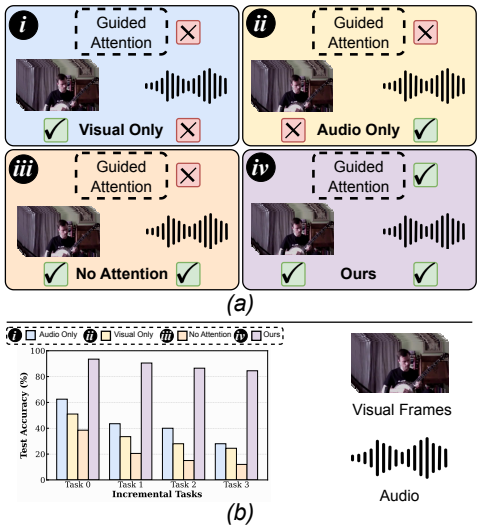

Figure 1. Overview of our proposed setting on the AVE dataset. (a) We present the four approaches: *i* only visual modality; *ii* only audio modality; *iii* both modalities fused naively (baseline); *iv* our proposed approach with guided attention-based fusion. (b) The superior performance of our method over individual modalities and the baseline underscores the necessity of coupling multimodal representations with guided attention for robust class-incremental learning.

1. Introduction

Class-Incremental Learning (CIL) aims to learn new classes incrementally without forgetting the previously learned ones. Prior works have leveraged different strategies including knowledge distillation (Castro et al., 2018; Dhar et al., 2019; Douillard et al., 2020), adapter-based (He et al., 2025; Wang et al., 2025), data/memory-replay (Qi et al., 2025; Channappayya et al., 2023; Chen et al., 2023), etc. The CIL setting is widely used across diverse practical applications, including robotics (Yao et al., 2025), autonomous driving, and medical imaging (Yi et al., 2025), where learnable systems must adapt to new objects or diseases without retraining from scratch. Despite significant advances, understanding class-incremental learning in foundational multimodal learning remains limited.

Foundational models (Radford et al., 2021; Achiam et al., 2023; Abdin et al., 2024; Kirillov et al., 2023) are large-scale models pre-trained on vast, diverse datasets to act as a general-purpose base for various downstream applications. They have shown great advancements in low-level vision tasks (Yadav et al., 2025b), virtual assistants (Kumar et al., 2025), autonomous driving (Guan et al., 2024), medical imaging (Azad et al., 2023), etc. Multimodal learning focuses on targeting two or more different modalities, from language, vision, and audio, to interact and provide a better understanding of the scene. Recent works in vision-language continual models (Jha et al., 2024; Cao et al., 2024) have focused on recognizing new classes while retaining strong, pre-trained capabilities. Similarly, in audio-visual modalities, some methods (Pian et al., 2023; 2024; Hong

This work was done by Avi Gupta during his internship at Dolby Laboratories. [1]IIIT Delhi [2]Dolby Laboratories [3]IIT Kanpur [4]TYUST China. Correspondence to: Avi Gupta <avig@iiitd.ac.in>.

Presented at the ICML 2026 Workshop "Continual Adaptation at Scale: Towards Sustainable AI".

et al., 2025) target to fuse both audio and visual feature representations to understand the scene in an incremental setting. However, these representations are extracted from separate encoders and require an additional strategy to align them in a common space. To address that, we propose leveraging SAM-Audio (Shi et al., 2025), a foundational multimodal model that already has this innate ability due to its pre-training.

To assess the robustness of SAM-Audio's multimodal representational space in the incremental setting, we conduct a preliminary empirical analysis of its audio and visual features. We compare four distinct configurations, as illustrated in Fig. 1(a): ***i*** unimodal visual features; ***ii*** unimodal audio features; ***iii*** a baseline utilizing naive audio-visual fusion; ***iv*** our proposed guided attention-based fusion; The empirical results in Fig. 1(b) reveal that relying on isolated unimodal representations yields severe performance degradation across successive incremental tasks. Crucially, the naive fusion baseline fails to mitigate catastrophic forgetting, resulting in suboptimal overall accuracy. These findings demonstrate that simply leveraging individual modalities or employing naive fusion strategies are insufficient for CIL.

To address these limitations, we introduce a guided strategy designed to leverage SAM-Audio's frozen representations for CIL. While large multimodal foundation models encapsulate rich audio-visual priors, our approach effectively leverages these static latent spaces to sequential learning tasks. Specifically, we extract features from the frozen SAM-Audio's audio and visual encoders and process them through a guided attention mechanism. In this module, audio representations serve as a contextual signal to guide the visual features toward precise target predictions. To further mitigate catastrophic forgetting, we employ strict knowledge distillation objectives at both the feature and logit levels. This ensures that the topology of previously learned multimodal alignments is preserved as new classes are introduced. As shown in Fig. 1(b), our proposed approach consistently outperforms the baselines across all incremental tasks. We summarize our key contributions as follows: ❶ We propose a novel, efficient paradigm for adapting large-scale, pretrained audio-visual foundation models (specifically SAM-Audio) to the challenging class-incremental learning setting, bypassing the need for computationally expensive full fine-tuning. ❷ We introduce a specialized guided attention mechanism that leverages dense audio representations as a conditioning signal to dynamically adapt visual features for sequential semantic prediction. ❸ To explicitly mitigate catastrophic forgetting, we formulate a robust distillation objective operating at both the feature and logit levels, ensuring the preservation of previously learned classes as new classes are introduced. ❹ Extensive experiments demonstrate that our proposed framework significantly outperforms existing state-of-the-art methods across standard benchmarks.

2. Proposed Method

2.1. Overview

The objective is to leverage the dense representations of the foundational multimodal model, SAM-Audio, and strategically adapt to new, unknown classes in an incremental setting while avoiding forgetting previously learned ones. The incremental setting is structured as a sequence of tasks $t \in \{0, 1, ..., T\}$. Each task t provides a distinct set of classes C_t. Following (Pian et al., 2023), the training dataset for the task t includes $\mathcal{D}'_t = \mathcal{D}_t \cup \mathcal{M}_{t-1}$, where $\mathcal{D}_t$ includes all the samples from current task while $\mathcal{M}_{t-1}$ comprises exemplar samples from the previous task. These exemplar samples are randomly selected from all tasks up to $t-1$. Each sample consists of a video sample x and its corresponding label y.

Our proposed approach is divided into three major components: *(a) Multimodal Feature Extractor*, that extracts the audio-visual features from multimodal encoder (Shi et al., 2025); *(b) Guided Attention Mechanism*, that fuses the extracted audio and visual features using transformer-based attention and provides sample-specific multimodal composite features; and *(c) Classifier*, that finally predicts the logits from the multimodal features. The overview of our proposed architecture is visualized in Fig. 2.

2.2. SAM-Audio as the Backbone

As mentioned before, the objective of our proposed approach is to leverage the learned representations of SAM-Audio for the incremental setting. Hence, we use the SAM-Audio as the backbone for both audio and visual modality. The video encoder of the SAM-Audio is trained on over 100 million videos. We also leverage the SAM-Audio's audio encoder. The audio encoder is based on the DAC-like autoencoder structure with a Variational Autoencoder (VAE) bottleneck layer. The audio is encoded into a sequence at 25 Hz. Similar to the video encoder, we keep the weight of the pretrained audio encoder frozen as well.

Given an input video sample x, we extract the video frames and audio signal ($x = \{x_v, x_a\}$) where, $x_v \in \mathbb{R}^{T_v \times C \times H \times W}$ and audio signal $x_a \in \mathbb{R}^k$. We use the visual ψ_v and audio ψ_a encoder to generate the corresponding visual features $f_v \in \mathbb{R}^{T_v \times d}$ and audio features $f_a \in \mathbb{R}^{T_a \times d}$ as: $f_v = \psi_v(x_v)$; and $f_a = \psi_a(x_a)$, respectively.

Here, T_a and T_v are the audio chunks and visual frames, respectively. C, H, and W are the number of channels, the height, and the width of each frame, respectively. d represents the hidden dimension. These extracted features are passed through the transformer-based attention mechanism to leverage the audio representation for video classification.

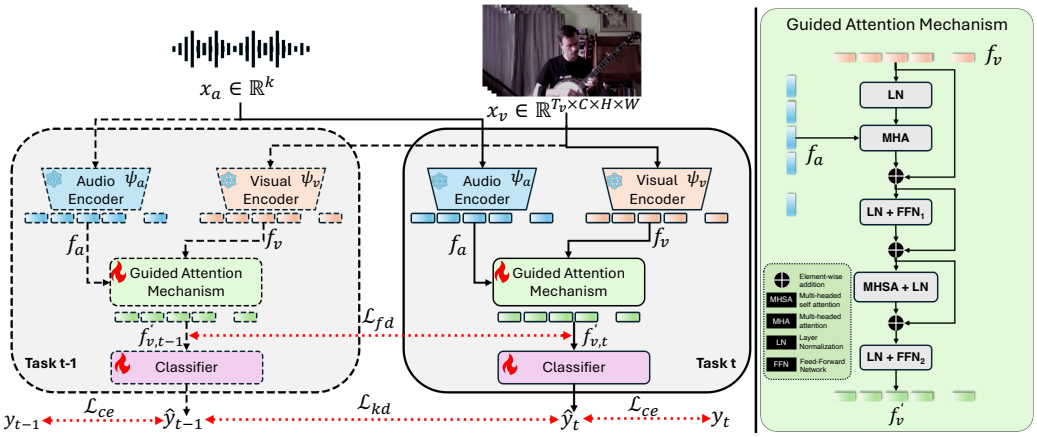

Figure 2. **[Left]** The overall architecture. At current task t, audio (f_a) and visual (f_v) representations are derived from audio signal (x_a) and video frames (x_v) using frozen audio (ψ_a) and visual (ψ_v) encoders, respectively of the SAM-Audio. These audio and frame embeddings are fused by a guided attention mechanism. The final video representations (f'_v) are passed through the classifier for precise class prediction. The model is trained using $\mathcal{L}_{ce}$ (for class prediction), $\mathcal{L}_{fd}$ (distillation at feature level), and $\mathcal{L}_{kd}$ (distillation at logits level). **[Right]** The guided attention mechanism, comprising a combination of multi-headed cross attention and multi-headed self attention, takes the audio representations and provides spatial and semantic guidance to the visual representation for precise class prediction.

2.3. Guided Attention Mechanism

The audio and visual embeddings from multimodal encoders provide robust and aligned information about the corresponding inputs. To effectively integrate these embeddings, inspired by (Jeong et al., 2025), we leverage a guided attention mechanism. This transformer-based attention provides effective guidance from audio, thus enhancing the visual content. The guided attention, visualized in Fig. 2[Right], consists of an attention block with a Multi-Head Attention (MHA) and a Feed-Forward Network (FFN) to blend the audio and frame representations. This is followed by Multi-Head Self-Attention (MHSA) and a Feed-Forward Network (FFN) for the representation refinement.

In particular, the attention block takes the audio f_a and frame f_v representations as input and fuses them together by MHA. This process is formulated as:

$$\begin{aligned} z &= \text{MHA}(\text{LN}(f_v), \text{LN}(f_a)) + f_v; \\ \overline{z} &= \text{FFN}_1(\text{LN}(z)) + z \end{aligned} \tag{1}$$

where LN denotes the layer normalization. The output representation $\overline{z}$ is refined to enhance overall visual quality. Specifically, $\overline{z}$ is passed through the Multi-Head Self-Attention (MHSA) module followed by another FFN to produce the refined visual representation f'_v. The refining process is formulated as:

$$\begin{aligned} \tilde{z} &= \text{MHSA}(\text{LN}(\overline{z})) + \overline{z}; \\ f'_v &= \text{FFN}_2(\text{LN}(\tilde{z})) + \tilde{z} \end{aligned} \tag{2}$$

2.4. Objective Function

Our proposed strategy employs *task-separated cross-entropy* $\mathcal{L}_{ce}$ for the class-incremental learning, which also prevents the prediction of old classes from being updated by learning new classes. The loss is formulated as:

$$\mathcal{L}_{ce} = \sum_{i=1}^{\mathcal{C}_t} y_i \log(\hat{y}_i) \tag{3}$$

where $\hat{y}_i$, and y_i denotes the predicted and ground-truth labels respectively. Also, we use the *feature distillation loss* $\mathcal{L}_{fd}$ between the attended visual features of the current and the previous task (on the exemplar samples $\mathcal{M}_{t-1}$) to teach $f'_{v,t}$ to mimic $f'_{v,t-1}$. It is formulated as:

$$\mathcal{L}_{fd} = ||f'_{v,t} - f'_{v,t-1}||^2 \tag{4}$$

The last part of our final objective function is the *task-wise knowledge distillation* $\mathcal{L}_{kd}$, where we evaluate the Kullback-Leibler (KL) divergence between the logits of classes in the current ($\hat{y}_t^{C_s}$) and previous ($\hat{y}_{t-1}^{C_s}$) tasks which helps preserve past knowledge and prevent the learned knowledge from being biased by other tasks.

$$\mathcal{L}_{kd} = \sum_{s=1}^{t} KL(\hat{y}_t^{C_s} || \hat{y}_{t-1}^{C_s}) \tag{5}$$

To sum it up, our overall ***objective function*** is formulated as: $\mathcal{L}_{tot} = \mathcal{L}_{ce} + \mathcal{L}_{fd} + \mathcal{L}_{kd}$

3. Experiments

3.1. Experimental Details

Datasets: For fair comparison, following (Pian et al., 2023), we evaluate our proposed approach on AVE-Class-Incremental (AVE-CI) dataset which comprises 4K+ videos covering 28 audio-visual event classes that are uniformly

distributed into 4 incremental tasks, with 7 classes each; and VGGSound100-Class-Incremental (VS100-CI) dataset with 100 classes and 60K samples that are distributed into 10 incremental tasks, with 10 classes in each.

Evaluation Metrics: We evaluate the performance of our approach with Mean Accuracy:

$$MeanAcc = \frac{1}{T}\sum_{t=1}^{T} a_t \tag{6}$$

where a_t is the testing accuracy of all seen classes after completing the training on the current task t.

We also evaluate the average forgetting to measure the extent of catastrophic forgetting over previously learned tasks:

$$\begin{aligned} AvgForgetting &= \frac{1}{T-1}\sum_{t=2}^{T} F_t \\ F_t &= \frac{1}{t-1}\sum_{i=1}^{t-1} \max_{k\in\{i,\ldots,t-1\}} (a_{k,i} - a_{t,i}) \end{aligned} \tag{7}$$

where $a_{k,i}$ is the testing accuracy of the i-th task after training on the k-th task.

Implementation Details: We leverage the foundational SAM-Audio for the audio and visual encoder. We freeze the pre-trained visual and audio encoders during training and fine-tune the rest of the model (audio-guided attention and the classifier). We use the SGD optimizer with a learning rate of 5e-5 and a weight decay of 1e-4, and use Cosine Annealing for the learning rate decay. For the AVE-CI and VS100-CI datasets, we set the maximum number of training epochs per incremental step to 200.

Table 1. Comparison of our proposed approach with current state-of-the-art methods. **Bold** denotes the best results.

Methods	AVE-CI		VS100-CI	
	Mean Acc.↑	Avg. Forgetting↓	Mean Acc.↑	Avg. Forgetting↓
Fine-tuning	42.40	70.99	26.21	89.37
LwF (Li & Hoiem, 2018)	58.07	26.90	59.34	23.01
iCaRL-NME (Rebuffi et al., 2017)	56.15	11.71	56.19	12.80
iCaRL-FC (Rebuffi et al., 2017)	65.88	26.08	64.22	29.94
SS-IL (Ahn et al., 2021)	61.94	22.49	69.20	9.75
AFC-NME (Kang et al., 2022)	68.46	14.18	61.41	23.30
AFC-LSC (Kang et al., 2022)	65.21	28.11	57.76	29.64
AV-CIL (Pian et al., 2023)	74.04	7.63	72.80	5.49
Ours	**88.72**	**2.14**	**79.64**	**0.34**

3.2. Results and Analysis:

In this section, we compare our proposed approach for audio-visual CIL with the current state of the art. The results are illustrated in Table 1. From the table, we observe that our proposed approach surpasses the previous works by a significant margin. Particularly, on the AVE-CI dataset, our approach consistently outperforms previous state-of-the-art (Pian et al., 2023) in both mean test accuracy and average forgetting. Similar performance gains are observed on VS100-CI, highlighting the robustness of our method across benchmarks.

3.3. Ablation Studies

Effectiveness of loss functions. In Table 2, we demonstrate the effectiveness of our individual loss functions on the AVE-CI dataset. We observe that removing both the feature ($\mathcal{L}_{fd}$) and knowledge distillation ($\mathcal{L}_{kd}$) losses results in poor test accuracy and further increases the risk of forgetting. To identify the individual contribution of both distillation-based losses, we remove $\mathcal{L}_{fd}$ and $\mathcal{L}_{kd}$ separately. We observe that including only $\mathcal{L}_{kd}$ improves performance, and including $\mathcal{L}_{fd}$ demonstrates *negative forgetting*, indicating that the model does not actually forget prior knowledge. Hence, both losses play an important role in enhancing overall performance and mitigating catastrophic forgetting.

Table 2. Ablation study to analyze the effectiveness of loss components and guided attention. **Bold** denotes the best results.

Guided Attention	$\mathcal{L}_{ce}$	$\mathcal{L}_{kd}$	$\mathcal{L}_{fd}$	Mean Acc.↑	Avg. Forgetting↓
✗	✓	✓	✗	21.45	12.58
✓	✓	✗	✗	87.26	4.30
✓	✓	✓	✗	87.34	2.29
✓	✓	✗	✓	87.79	**-1.32**
✓	✓	✓	✓	**88.72**	2.14

Table 3. Ablation study to analyze the effectiveness of individual modality on AVE dataset. **Bold** denotes the best results.

Modality		Mean Acc.↑	Avg. Forgetting↓
Audio	Video		
✓	✗	34.21	11.30
✗	✓	43.50	5.66
✓	✓	**88.72**	**2.14**

Effectiveness of individual modality. In Table 3, we perform another study to analyze the effectiveness of both audio and visual modalities in the class-incremental setting. We observe that using both modalities significantly improves performance and reduces forgetting compared to using either modality alone. Furthermore, we see some improvement when using only the visual modality over audio, both in accuracy and in forgetting; however, the individual modalities do not keep up with our approach. The per-step test accuracy of individual audio and visual modality is also visualized in Fig. 1(b). Additional analysis and ablation studies are provided in the ***Appendix***.

4. Conclusion

In this work, we explore the task of continual learning using foundational multimodal models to effectively learn new visual classes with audio guidance, without catastrophically forgetting previously learned knowledge. To address the proposed problem, we leverage the dense representations of the foundational multimodal model, SAM-Audio. Furthermore, we propose an audio-guided transformer-based attention mechanism. Finally, we utilize knowledge distillation at the task and feature levels to retain prior knowledge. Experiments on benchmark datasets show that our approach outperforms the current state-of-the-art methods by a significant margin. Given the significant performance improvement from SAM-Audio, we plan to generalize our proposed approach to other foundational multimodal models as future work.

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Ahn, H., Cha, S., Lee, D., and Moon, T. Uncertainty-based continual learning with adaptive regularization. *Advances in neural information processing systems*, 32, 2019.

Ahn, H., Kwak, J., Lim, S., Bang, H., Kim, H., and Moon, T. Ss-il: Separated softmax for incremental learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 824–833, 2021. doi: 10.1109/ICCV48922.2021.00088.

Azad, B., Azad, R., Eskandari, S., Bozorgpour, A., Kazerouni, A., Rekik, I., and Merhof, D. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023.

Benjamin, A. S., Rolnick, D., and Kording, K. Measuring and regularizing networks in function space. *arXiv preprint arXiv:1805.08289*, 2018.

Cao, M., Liu, Y., Liu, Y., Wang, T., Dong, J., Ding, H., Zhang, X., Reid, I., and Liang, X. Continual llava: Continual instruction tuning in large vision-language models. *arXiv preprint arXiv:2411.02564*, 2024.

Castro, F. M., Marín-Jiménez, M. J., Guil, N., Schmid, C., and Alahari, K. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 233–248, 2018.

Channappayya, S., Tamma, B. R., et al. Augmented memory replay-based continual learning approaches for network intrusion detection. *Advances in Neural Information Processing Systems*, 36:17156–17169, 2023.

Chaudhry, A., Dokania, P. K., Ajanthan, T., and Torr, P. H. S. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018a.

Chaudhry, A., Ranzato, M., Rohrbach, M., and Elhoseiny, M. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018b.

Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P. K., Torr, P. H., and Ranzato, M. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Chen, J., Cong, R., Luo, Y., Ip, H., and Kwong, S. Saving 100x storage: Prototype replay for reconstructing training sample distribution in class-incremental semantic segmentation. *Advances in Neural Information Processing Systems*, 36:35988–35999, 2023.

Dhar, P., Singh, R. V., Peng, K.-C., Wu, Z., and Chellappa, R. Learning without memorizing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5138–5146, 2019.

Douillard, A., Cord, M., Ollion, C., Robert, T., and Valle, E. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European conference on computer vision*, pp. 86–102. Springer, 2020.

Golkar, S., Kagan, M., and Cho, K. Continual learning via neural pruning. *arXiv preprint arXiv:1903.04476*, 2019.

Guan, Y., Liao, H., Li, Z., Hu, J., Yuan, R., Zhang, G., and Xu, C. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*, 2024.

He, J., Duan, Z., and Zhu, F. Cl-lora: Continual low-rank adaptation for rehearsal-free class-incremental learning. *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 30534–30544, June 2025.

Hong, Y., Yang, Q., Zhang, T., Wang, Z., Fu, Z., Ding, K., Fan, B., and Xiang, S. Taming modality entanglement in continual audio-visual segmentation. *CoRR*, abs/2510.17234, 2025. doi: 10.48550/ARXIV.2510.17234. URL https://doi.org/10.48550/arXiv.2510.17234.

Hou, S., Pan, X., Loy, C. C., Wang, Z., and Lin, D. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 831–839, 2019.

Hung, C.-Y., Tu, C.-H., Wu, C.-E., Chen, C.-H., Chan, Y.-M., and Chen, C.-S. Compacting, picking and growing for unforgetting continual learning. *Advances in neural information processing systems*, 32, 2019.

Jeong, B., Park, J., Kim, S., and Kwak, S. Learning audio-guided video representation with gated attention for video-text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26202–26211, June 2025.

Jha, S., Gong, D., and Yao, L. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *Advances in neural information processing systems*, 37: 129146–129186, 2024.

Jung, S., Ahn, H., Cha, S., and Moon, T. Continual learning with node-importance based adaptive group sparse regularization. *Advances in neural information processing systems*, 33:3647–3658, 2020.

Kang, M., Park, J., and Han, B. Class-Incremental Learning by Knowledge Distillation with Adaptive Feature Consolidation. In *CVPR*, 2022.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Kumar, S., Nutalapati, P., Vemuri, S. S., Aida, R., Salami, Z. A., and Boob, N. S. Gpt-powered virtual assistants for intelligent cloud service management. In *2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS)*, pp. 1–6. IEEE, 2025.

Li, X., Zhou, Y., Wu, T., Socher, R., and Xiong, C. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International conference on machine learning*, pp. 3925–3934. PMLR, 2019.

Li, Z. and Hoiem, D. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. doi: 10.1109/TPAMI.2017.2773081.

Pian, W., Mo, S., Guo, Y., and Tian, Y. Audio-visual class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7799–7811, 2023.

Pian, W., Nan, Y., Deng, S., Mo, S., Guo, Y., and Tian, Y. Continual audio-visual sound separation. In Globersons, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J. M., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Qi, Z., Tang, Y.-P., Meng, L., Yu, H., Li, X., and Meng, X. Class-wise balancing data replay for federated class-incremental learning. *arXiv preprint arXiv:2507.07712*, 2025.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5533–5542, 2017. doi: 10.1109/CVPR.2017.587.

Shi, B., Tjandra, A., Hoffman, J., Wang, H., Wu, Y.-C., Gao, L., Richter, J., Le, M., Vyas, A., Chen, S., Feichtenhofer, C., Dollár, P., Hsu, W.-N., and Lee, A. Sam audio: Segment anything in audio. 2025. URL https://arxiv.org/abs/2512.18099.

Sun, Y., Si, Y., Zhu, C., Zhang, K., Shui, Z., Ding, B., Lin, T., and Yang, L. Cpathagent: An agent-based foundation model for interpretable high-resolution pathology image analysis mimicking pathologists' diagnostic logic. *arXiv preprint arXiv:2505.20510*, 2025.

Wang, F.-Y., Zhou, D.-W., Ye, H.-J., and Zhan, D.-C. Foster: Feature boosting and compression for class-incremental learning. In *European conference on computer vision*, pp. 398–414. Springer, 2022.

Wang, Y., Wang, L., Zhou, Q., Wang, Z., Li, H., Hua, G., and Tang, W. Multimodal llm enhanced cross-lingual cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8296–8305, 2024.

Wang, Y., Zhou, D.-W., and Ye, H.-J. Integrating task-specific and universal adapters for pre-trained model-based class-incremental learning. In *ICCV*, 2025.

Wu, H., Zhang, Z., Zhang, E., Chen, C., Liao, L., Wang, A., Xu, K., Li, C., Hou, J., Zhai, G., Xue, G., Sun, W., Yan, Q., and Lin, W. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25490–25500, June 2024.

Yadav, S., Gupta, A., and Jerripothula, K. R. Samwave: Adapting segment anything model to difficult tasks. In *36th British Machine Vision Conference 2025, BMVC 2025, Sheffield, UK, November 24-27, 2025*. BMVA, 2025a. URL https://bmva-archive.org.uk/bmvc/2025/assets/papers/Paper_698/paper.pdf.

Yadav, S., Gupta, A., and Jerripothula, K. R. Samwave: Wavelet-driven feature enrichment for effective adaptation of segment anything model. *arXiv preprint arXiv:2507.20186*, 2025b.

Yao, Y., Liu, S., Song, H., Qu, D., Chen, Q., Ding, Y., Zhao, B., Wang, Z., Li, X., and Wang, D. Think small, act big:

Primitive prompt learning for lifelong robot manipulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22573–22583, 2025.

Yi, H., Xu, W., Qin, Z., Chen, X., Wu, X., Li, K., and Lao, Q. idpa: Instance decoupled prompt attention for incremental medical object detection. *arXiv preprint arXiv:2506.00406*, 2025.

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm Ai*, 2(1):AIoa2400640, 2025.

Zhang, Z., Wu, H., Zhang, E., Zhai, G., and Lin, W. Q-bench$^+$+: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10404–10418, 2024. doi: 10.1109/TPAMI.2024.3445770.

A. Related Works

Class-Incremental Learning. The task of Class-Incremental Learning (CIL) targets to learn new classes continuously while preserving knowledge of previously learned classes (Jung et al., 2020; Hou et al., 2019; Kang et al., 2022; Wang et al., 2022). Several strategies have been adopted to overcome catastrophic forgetting and retain previously learned knowledge. Parameter regularization-based methods (Ahn et al., 2019; Benjamin et al., 2018; Chaudhry et al., 2018a) aim to estimate the importance of different parameters of the model and allocate them with different weights to indicate their importance. Knowledge distillation-based methods (Castro et al., 2018; Dhar et al., 2019; Douillard et al., 2020) help the model preserve previously learned knowledge across incremental steps by minimizing the Kullback-Leibler divergence between the output probability distributions of the previous and current models. Exemplar/Memory Replay-based methods (Ahn et al., 2021; Chaudhry et al., 2018b; 2019; Channappayya et al., 2023; Chen et al., 2023) assume that a small size of memory is accessible to store examples from old tasks/classes. Architecture-based methods (Golkar et al., 2019; Hung et al., 2019; Li et al., 2019) hold incremental modules to increase the capacity of the model to handle new tasks/classes.

Foundational Multimodal Learning. Multimodal foundation models (Radford et al., 2021; Achiam et al., 2023; Abdin et al., 2024) have demonstrated remarkable efficacy across diverse applications—ranging from low-level vision tasks (Yadav et al., 2025a; Zhang et al., 2024; Wu et al., 2024) and image recognition (Zhang et al., 2025) to cross-modal retrieval (Wang et al., 2024) and agent-based reasoning (Sun et al., 2025)—by aligning and binding different modalities within a joint embedding space. Benefiting from large-scale pretraining, these architectures excel at capturing complex cross-modal semantic associations. A prominent example is the Segment Anything Model for Audio (SAM-Audio) (Shi et al., 2025), which leverages advanced latent spaces to unify audio, visual, and textual modalities for prompt-driven sound separation. However, while such models exhibit strong generalization capabilities, they face severe limitations in dynamic environments that necessitate continual adaptation. Specifically, in the challenging setting of CIL, these models remain highly susceptible to catastrophic forgetting (shown in Fig 1). Our objective is to harness the robust multimodal priors embedded in SAM-Audio and adapt its architecture to enable scalable, resilient class-incremental learning.

B. Additional Results and Analysis

In Fig. 3, we comparatively analyze the test accuracy of the proposed approach with AV-CIL (Pian et al., 2023) on AVE-CI and VS100-CI at individual incremental steps. From the plots, we observe that our proposed approach consistently outperforms AV-CIL across all the incremental tasks in both benchmark datasets.

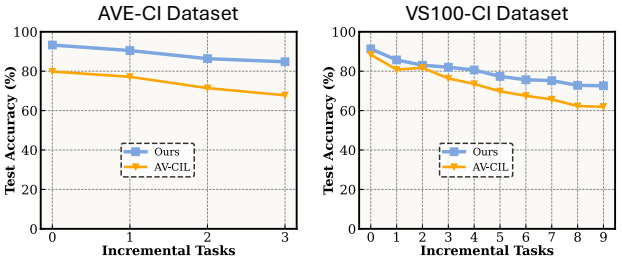

Figure 3. Performance (test accuracy) comparison with AV-CIL (Pian et al., 2023). **[Left]** Individual task performance for AVE-CI dataset on 4 incremental tasks. **[Right]** Individual task performance for VS100-CI dataset on 10 incremental tasks.

To further assess the robustness of our proposed approach, we perform an additional experiment on different class-incremental tasks. Following (Pian et al., 2023), we provide a new task setting with 4 distinct classes in each task, making a total of 7 tasks. This setting makes the problem more challenging with more sequential steps. As observed in Table 4, our proposed approach outperforms previous methods by a significant margin. Crucially, we observe a small drop in test accuracy compared to the traditional 4-step setting. We find that this drop is mainly due to an increase in the number of incremental steps.

C. Additional Ablation Studies

Importance of Guided Attention. While SAM-Audio provides a strong foundational model with highly dense representations, naively fusing its audio and visual modalities fails to adapt effectively in a class-incremental scenario. This simplistic fusion not only restricts overall accuracy but also triggers severe catastrophic forgetting as new tasks are introduced. The empirical vulnerability of naive fusion is demonstrated in Table 2 [row 1], which highlights a rapid deterioration of previously acquired concepts. To circumvent this, we employ our proposed guided attention alongside auxiliary feature-level distillation losses to preserve and integrate knowledge. The inclusion of guided attention significantly bolsters performance and stabilizes the model against forgetting [row 5]. The corresponding task-wise accuracy for this ablation is further visualized in Fig. 1(b).

Table 4. Comparison of our proposed approach with previous methods under different class-incremental settings on AVE-CI dataset. **Bold** denotes the best test accuracy (higher is better).

Settings	Fine-tuning	LwF (Li & Hoiem, 2018)	iCaRL-NME (Rebuffi et al., 2017)	iCaRL-FC (Rebuffi et al., 2017)	SS-IL (Ahn et al., 2021)	AFC-NME (Kang et al., 2022)	AFC-LSC (Kang et al., 2022)	AV-CIL (Pian et al., 2023)	Ours
7 classes × 4 steps	42.40	58.07	56.15	65.88	61.94	68.46	65.21	74.04	**88.72**
4 classes × 7 steps	37.54	50.25	60.02	65.50	65.29	68.61	61.93	71.76	**86.36**