Pseudo-Rehearsal for Audio-Visual Continual Learning on Embedded Systems

Jonathan Grienay*†‡, Martial Mermillod†, Laurent Rodriguez‡, Benoit Miramond‡ and Marina Reyboz*
*Univ. Grenoble Alpes, CEA, LIST, Grenoble, France
{jonathan.grienay, marina.reyboz}@cea.fr
†Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, Grenoble, France
martial.mermillod@univ-grenoble-alpes.fr
‡LEAT, Université Côte d'Azur / CNRS UMR 7248, Sophia Antipolis, France
{laurent.rodriguez, benoit.miramond}@univ-cotedazur.fr

***Abstract*—Continual learning for multimodal audio-visual systems on embedded devices faces dual challenges: catastrophic forgetting and privacy constraints prohibiting raw data storage. We present AV-DreamNet, the first noise-based pseudo-rehearsal method for audio-visual continual learning. By generating pseudo-samples from noise rather than storing real data, our approach enables privacy-preserving learning without auxiliary generative models or large buffers. On CL-VGGSound, AV-DreamNet achieves 47.1% (datafree version) and 58.1% (small buffers) accuracy versus 26.3% for state-of-the-art buffer-based methods. Considering embedded deployment, working in latent space enables offloading frozen encoders to dedicated inference accelerators and avoids heavy decoder architectures, while latent-space buffers provide $\sim 360\times$ memory reduction compared to raw data storage.**

***Index Terms*—continual learning, pseudo-rehearsal, audio-visual learning, catastrophic forgetting, embedded systems**

I. Introduction

The human brain seamlessly integrates audio-visual information across time, continually learning from new experiences without forgetting previously acquired knowledge. This capability contrasts sharply with current artificial neural networks, which suffer from catastrophic forgetting [1], [2]: when learning new information, they dramatically lose performance on previously learned tasks. As intelligent systems proliferate in resource-constrained environments—smart homes, wearable devices, autonomous robots—the ability to learn continually from multimodal streams becomes increasingly critical, yet challenging to implement on embedded hardware due to learning computation overhead.

The problem is compounded by three interconnected challenges. First, neural networks exhibit catastrophic forgetting. Second, computational limitations of edge devices preclude the deployment of large pretrained models that have become dominant in multimodal learning [3]. Third, privacy regulations (GDPR, healthcare standards) and personal device constraints prohibit storing raw user data for replay-based continual learning methods [4].

This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the 'PEPR IA France 2030' programme (Emergences project ANR-23-PEIA-0002)

Among continual learning strategies, replay-based methods have been widely explored but most rely on large buffers containing real data leading to privacy leaks. Pseudo-rehearsal [4] uniquely addresses privacy by generating synthetic samples from noise, without storing real data or training auxiliary generative models. However, it has only been applied to unimodal visual tasks. Meanwhile, recent multimodal continual learning methods [3], [5], [6] show that audio-visual information reduces forgetting, but all rely on buffer-based replay, requiring raw sample storage from prior tasks.

The recent CGIL method [7] represents the first application of generative replay to multimodal continual learning, using VAE-based latent replay for vision-language tasks with CLIP. While promising, CGIL's reliance on large pretrained models (approximately 150M parameters) and VAE training makes it very heavy for embedded deployment. Moreover, VAE-based generation can produce samples closely resembling real training data, potentially compromising privacy guarantees. Furthermore, no prior work has addressed audio-visual continual learning through pseudo-rehearsal, leaving a critical gap for applications requiring privacy-preserving multimodal learning on edge devices.

In this work, we introduce the first noise-based pseudo-rehearsal method for audio-visual continual learning. By adapting pseudo-rehearsal [4] to multimodal inputs, our method preserves learned knowledge without storing raw data or training separated generative models. Extending our pseudo-rehearsal mechanisms with small buffers (buffer-seeded replay approach, [8]), we show that minimal real samples as seeds enhance pseudo-sample quality. Our embedded-friendly architecture uses frozen lightweight encoders, reducing continual training cost and memory footprint.

Experiments on CL-VGGSound [3] demonstrate that our approach achieves competitive performance with buffer-based methods while requiring significantly less memory and computational resources. We provide comprehensive comparisons with recent audio-visual continual learning baselines [3], establishing the viability of pseudo-rehearsal generative replay for practical privacy-preserving multimodal continual learning on embedded systems.

II. Related Work

A. Continual Learning and Catastrophic Forgetting

Catastrophic forgetting [1] remains a fundamental challenge in neural networks, where learning new tasks causes dramatic performance degradation on previously learned tasks. Three main paradigms address this problem: regularization-based, replay-based, and architecture-based approaches. Replay methods can be further subdivided into buffer-based, generative, and pseudo-rehearsal approaches.

Regularization-based methods constrain weight updates to preserve previous knowledge, using Fisher information (EWC [2]), online importance tracking (SI [9]), or knowledge distillation (LwF [10]). While computationally efficient, these methods typically exhibit limited performance on long task sequences.

Architecture-based methods dynamically expand network capacity for new tasks. DER [11] freezes previous representations and augments with new dimensions, while FOSTER [12] uses feature boosting and compression. These methods face scalability challenges or incur computational overhead from distillation.

Replay-based methods reintroduce exemplars from previous tasks for rehearsal during new task learning. iCaRL [13] combines exemplar storage with distillation, using herding to select representative samples. Dark Experience Replay (DER++) [14] stores both exemplars and their logits, achieving strong performance across various continual learning scenarios. However, replay methods require memory proportional to the number of tasks and raise privacy concerns when storing sensitive data.

Generative replay addresses privacy and memory limitations by synthesizing realistic samples instead of storing real data. Deep Generative Replay [15] pioneered this approach using GANs, training a generator alongside the task solver. Brain-Inspired Replay [16] introduced five brain-inspired components including context-modulated feedback and internal replay, achieving substantial improvements over standard generative replay (17.3 percentage points on Split CIFAR-100). Recent advances include diffusion-based generative replay (DDGR) [17], which produces higher-quality samples than GAN-based methods, and latent generative replay [18], which generates in compressed latent spaces for resource efficiency. However, these methods require training and maintaining separate generative models, adding computational and memory overhead.

Pseudo-rehearsal offers a lightweight alternative by capturing the network's transfer function without explicit generative models. DreamNet [4], [19] employs noise-based pseudo-rehearsal: by injecting random noise and capturing the network's input-output mappings, this approach can be viewed as implicit self-distillation that preserves learned knowledge without storing real data. A critical enhancement is the *reinjection* mechanism [20]: pseudo-samples are iteratively fed back through the network's auto-associative output, progressively refining them toward stable attractors that better represent the learned function. This iterative sampling significantly improves pseudo-sample quality without requiring additional generative models. Combined replay [8] further demonstrates that minimal real samples (as few as 1-5 per class) used as seeds can enhance pseudo-sample quality, outperforming pure experience replay with buffers smaller than 100 samples per class. However, all these methods have been applied exclusively to unimodal visual tasks, leaving multimodal pseudo-rehearsal unexplored for audio-visual domains.

B. Multimodal Continual Learning

Recent work has demonstrated that multimodal learning provides complementary information that can mitigate catastrophic forgetting. SAMM (Semantic-Aware Multimodal Method) [3] introduced the CL-VGGSound benchmark with three continual learning scenarios: class-incremental (Seq-VGGSound), domain-incremental (Dom-VGGSound), and generalized class-incremental learning. Using FiLM fusion and semantic-aware feature alignment, SAMM demonstrated that multimodal learning reduces error rates by 3-5x compared to single modalities, achieving 34.51% accuracy on Seq-VGGSound with 1000-sample buffers compared to 28.09% for standard experience replay.

CIGN (Class-Incremental Grouping Network) [5] introduces learnable audio-visual class tokens with audio-visual grouping mechanisms. The method achieves continual learning through class token distillation and continual grouping, demonstrating state-of-the-art performance on VGGSound-Instruments and VGGSound-100 benchmarks. However, CIGN requires memory buffers proportional to the number of classes and tasks.

ContAV-Sep [6] addresses continual audio-visual sound separation, proposing Cross-modal Similarity Distillation Constraint (CrossSDC) to preserve cross-modal semantic similarity across tasks. While demonstrating effective mitigation of catastrophic forgetting in sound separation, the method relies on distillation rather than generative mechanisms.

MMAL [21] proposes an exemplar-free approach using analytic learning with Recursive Least Squares, achieving 76.71%, 78.98%, and 76.19% on AVE, Kinetics-Sounds, and VGGSound-100 respectively. However, MMAL employs fundamentally different optimization (closed-form solutions) rather than generative replay.

IEMF [22] introduces inverse effectiveness principles from neuroscience, demonstrating that weaker unimodal cues lead to stronger fusion benefits. The method achieves up to 50% computational cost reduction while improving performance on audio-visual continual learning tasks. Despite its brain-inspired motivation, IEMF uses buffer-based replay rather than generative approaches.

CGIL [7] represents the most closely related work, being the first to combine generative replay with multimodal (vision-language) continual learning. CGIL uses VAEs to learn class-conditioned distributions in CLIP's visual embedding space, generating synthetic features for prompt tuning. While achieving state-of-the-art performance on vision-language benchmarks, CGIL differs fundamentally from our work: (1) it

targets vision-language rather than audio-visual domains, (2) requires large pretrained models (CLIP with approximately 150M parameters), (3) uses VAE-based latent replay requiring additional generative model training, and (4) is designed for general continual learning rather than embedded deployment. Our work addresses the distinct challenge of audio-visual continual learning with lightweight, noise-based generation suitable for resource-constrained devices.

C. Research Gap and Positioning

Our literature review reveals a confirmed gap: while generative replay has been successfully applied to unimodal visual tasks (DreamNet, Brain-Inspired Replay) and recently extended to vision-language with large models (CGIL), no prior work combines pseudo-rehearsal with audio-visual continual learning for embedded systems. All existing audio-visual continual learning methods (SAMM, CIGN, ContAV-Sep, IEMF) rely on buffer-based replay or distillation, raising privacy concerns when handling sensitive audio-visual data.

Our work addresses this gap by: (1) extending DreamNet's noise-based pseudo-rehearsal to audio-visual inputs, (2) demonstrating embedded-friendly multimodal continual learning with 7x fewer parameters than CLIP-based approaches, and (3) providing the first comprehensive evaluation of generative replay for privacy-preserving audio-visual continual learning on resource-constrained devices.

III. Method

We present AV-DreamNet for extending DreamNet [4] pseudo-rehearsal to audio-visual continual learning, exploring different fusion strategies. The core principle is generating pseudo-samples from noise to rehearse previous knowledge without storing or regenerate real data.

A. Problem Formulation

In continual audio-visual learning, we encounter a sequence of tasks $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$ arriving sequentially. Each task T_t consists of audio-visual-label tuples (a, v, y) where $a \in \mathbb{R}^S$ represents raw audio waveforms (S samples), $v \in \mathbb{R}^{H\times W\times 3}$ represents RGB video frames, and $y \in \{1, \ldots, C_t\}$ denotes class labels. The goal is to learn a function $f_\theta : (\mathcal{A}, \mathcal{V}) \to \mathcal{Y}$ that performs well on all tasks seen so far, without catastrophic forgetting of previous tasks and without storing previous task data. We evaluate performance using standard continual learning metrics Average Accuracy AA $= \frac{1}{N}\sum_{i=1}^{N} acc_{N,i}$ where $acc_{N,i}$ is accuracy on task i after learning all N tasks).

B. Multimodal Pseudo-Rehearsal Mechanism

As illustrated in Figure 1a, to generate pseudo-samples for previous tasks, we maintain a frozen copy f_{t-1} of the model trained on the previous task. We sample noise vectors $X_A^{(0)}$ and $X_A^{(0)}$ as initial features (see equation 1) for audio and visual modalities respectively, then pass them through f_{t-1} to generate pseudo-labels and auto-associative reconstructions. This frozen model captures the transfer function learned up to task $t-1$, ensuring that pseudo-samples represent previously acquired knowledge without interference from current task learning. During training on task t, these synthetic audio-visual pairs are mixed with real samples using a replay ratio r: for each real sample, we generate $r \times K$ pseudo-samples, where K is the number of reinjections.

A key mechanism for improving pseudo-sample quality is iterative reinjection sampling [20]. Starting from noise, we feed the input through the network and collect the auto-associative outputs X'_A and X'_V. These reconstructed features are then *reinjected* as new inputs for subsequent iterations:

$$\begin{aligned} X_A'^{(k)}, X_V'^{(k)}, Y'^{(k)} &= f_{\text{t-1}}(X_A^{(k)}, X_V^{(k)}) \\ \text{with:} \quad & X_A^{(k)}, X_V^{(k)} = X_A'^{(k-1)}, X_V'^{(k-1)} \\ & (X_A^{(0)}, X_V^{(0)}) \sim \mathcal{N}(0, \sigma^2 I) \\ & k = 0, \ldots, K-1 \end{aligned} \tag{1}$$

Through this iterative process, pseudo-samples progressively converge toward stable attractor states that better capture the network's learned transfer function. The number of reinjection steps K controls the quality-computation trade-off: more reinjections produce better pseudo-samples but increase training time.

For each task t, the model is trained for N epochs on $\mathcal{D}_{\text{train}}$. Following DreamNet [4], we use Binary Cross-Entropy (BCE) for both classification and auto-associative reconstruction, which stabilizes reinjection-based pseudo-rehearsal:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(y, \hat{y}) + \lambda_A \mathcal{L}_{\text{BCE}}(X_A, X'_A) + \lambda_V \mathcal{L}_{\text{BCE}}(X_V, X'_V) \tag{2}$$

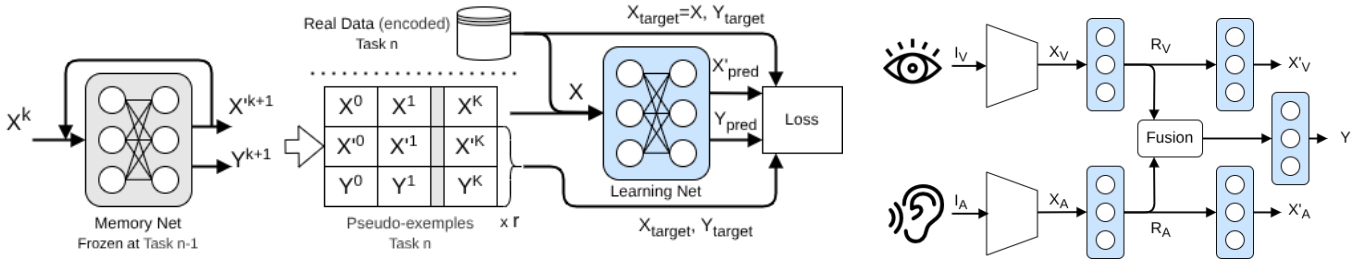
where $\mathcal{L}_{\text{BCE}}(y, \hat{y})$ is the classification loss, and $\mathcal{L}_{\text{BCE}}(X_A, X'_A)$, $\mathcal{L}_{\text{BCE}}(X_V, X'_V)$ are auto-associative reconstruction losses with modality-specific weights λ_A and λ_V. The use of BCE, rather than mean squared error, is crucial: BCE treats each feature dimension independently with bounded gradients, providing stable training when pseudo-samples are generated from noise, and preserves the probabilistic interpretation needed for reinjection-based refinement. The auto-associative losses encourage the representation to maintain information about both modalities, allowing for reinjections and preventing modality collapse.

C. Buffer-Seeded Replay Strategy

Following [8], [20], we implement buffer-seeded replay to enhance pseudo-sample quality. We maintain a minimal buffer of k samples per class as seeds. During training on task t, we randomly select buffer exemplars (a_i, v_i) as seed:

$$\begin{aligned} X_A^{(0)} &\sim a_i + \mathcal{N}(0, \sigma^2 I) \\ X_V^{(0)} &\sim v_i + \mathcal{N}(0, \sigma^2 I) \end{aligned} \tag{3}$$

This provides high-quality starting points for pseudo-sample generation. While storing even minimal real samples relaxes the strict privacy guarantees of pure pseudo-rehearsal, buffer-seeded replay offers a favorable trade-off: with just a few latent samples per class, we observe significant accuracy improvements over pure pseudo-rehearsal while using much



(a) Overview of the pseudo-rehearsal mechanism. The frozen Memory Net from task $n-1$ generates pseudo-samples from noisy inputs X^0 through iterative reinjection, producing refined features $(X'^{0\to K}, Y^{0\to K})$. These pseudo-samples are mixed with real data from task n and used to train the Learning Net.

(b) AV-DreamNet architecture for audio-visual continual learning. Each modality has its own auto-encoder pathway and fusion applied only for classification Y.

Fig. 1: Pseudo-rehearsal learning mechanism and AV-DreamNet architecture for audio-visual continual learning.

less memory than standard experience replay thanks to latent-space storage.

D. AV-DreamNet Architecture

We propose the AV-DreamNet architecture where each modality maintains its own complete auto-encoder pathway. Each auto-decoders are connected to unimodal representations and fused representation are only used for classification. We systematically compare three fusion strategies (concatenation, average, FiLM).

Figure 1b illustrates the AV-DreamNet architecture.

1) Fusion Strategies: The classification output Y is produced from a fused representation R_{AV} that combines both modalities. We investigate three fusion mechanisms:

- **Concatenation**: $R_{AV} = [X_A; X_V]$
- **Average**: $R_{AV} = \frac{1}{2}(X_A + X_V)$
- **FiLM** [3]: $R_{AV} = \gamma(X_V) \odot X_A + \beta(X_V)$

2) Reconstruction from Unimodal Representations: In AV-DreamNet, audio and visual inputs are processed through separate encoders to produce X_A and X_V, then auto-associative reconstruction occurs from the unimodal representations directly: X'_A is decoded from X_A and X'_V from X_V, independently of fusion. Fusion is applied only to produce the classification output Y. This design provides:

- **Modality Independence**: Each auto-encoder captures its modality's transfer function separately
- **Privacy Advantage**: Enables selective pseudo-rehearsal (audio-only or visual-only)
- **Robustness**: Better handling of missing modalities during inference

IV. Experiments

We evaluate AV-DreamNet on the CL-VGGSound benchmark across three continual learning scenarios, comparing against state-of-the-art buffer-based methods. Our experiments investigate the impact of key hyperparameters including fusion strategies, model depth, reinjection steps, replay ratio, and buffer size.

A. Experimental Setup

1) Dataset and Benchmark: We evaluate on CL-VGGSound [3], a standard benchmark for audio-visual continual learning. The dataset contains 100 classes of audio-visual events, split into sequential tasks for continual learning evaluation. We follow the three scenarios defined by [3]:

- **Seq-VGGSound**: Class-incremental learning with 10 tasks of 10 classes each
- **Dom-VGGSound**: Domain-incremental learning where the 100 classes (10 Tasks of 10 classes) are grouped into 5 supercategories ("home", "sport", "music", "people" and "vehicles") used as classification targets. We note that the reported SAMM results may not follow supercategory-level classification for this scenario, which could partly explain the performance gap. This scenario evaluates the model's ability to leverage shared properties within domains.
- **GCIL-VGGSound**: Generalized class-incremental learning with imbalanced task sizes, where each task contains a variable number of classes following a long-tail distribution. This challenging scenario better reflects real-world deployment conditions by introducing new examples from classes already present in early tasks.

2) Baselines: We compare against baselines from the SAMM benchmark [3]:

- **Joint**: Upper bound trained on all tasks simultaneously
- **Finetuning**: Lower bound with sequential training without any forgetting mitigation
- **ER**: Experience Replay storing raw audio-visual samples in a memory buffer, replayed during training on new tasks
- **SAMM**: Semantic-Aware Multimodal Method combining FiLM-based audio-visual fusion with semantic-aware feature alignment to preserve cross-modal relationships

We report results for buffer sizes of 0, 5, and 20 samples per class to evaluate the accuracy-memory trade-off. For our architecture, we additionally report Joint (upper bound) and Finetuning (lower bound) computed on our architecture.

3) Feature extraction: We use pretrained ImageNet ResNet18 for visual features (11.2M parameters) and CNN10

TABLE I: Results on Seq-VGGSound (Class-Incremental Learning). A = Audio, V = Visual, AV = Audio-Visual accuracy.

Method	Buffer = 0			Buffer = 5			Buffer = 20		
	A	V	AV	A	V	AV	A	V	AV
SAMM [3]									
JOINT	53.47	34.52	58.21	–	–	–	–	–	–
Finetuning	7.60	6.37	8.24	–	–	–	–	–	–
ER	–	–	–	13.92	9.07	21.44	23.41	14.19	34.02
SAMM	–	–	–	23.61	8.90	26.34	32.20	11.60	37.76
AV-DreamNet *(Ours)*									
JOINT	60.35±0.54	49.98±0.30	69.69±0.31	–	–	–	–	–	–
Finetuning	13.46±0.45	8.85±0.13	14.41±0.43	–	–	–	–	–	–
Unimodal	44.74±0.57	32.33±0.61	–	48.75±0.55	44.16±0.54	–	47.00±0.73	41.89±0.46	–
Concat	–	–	**47.09±0.55**	–	–	57.54±0.70	–	–	**56.08±0.46**
Avg	–	–	43.99±3.59	–	–	**58.08±0.68**	–	–	55.79±0.73
FiLM	–	–	38.26±1.40	–	–	52.06±0.98	–	–	49.09±1.10

TABLE II: Results on Dom-VGGSound (Domain-Incremental Learning). A = Audio, V = Visual, AV = Audio-Visual accuracy.

Method	Buffer = 0			Buffer = 5			Buffer = 20		
	A	V	AV	A	V	AV	A	V	AV
SAMM [3]									
JOINT	57.48	42.87	61.66	–	–	–	–	–	–
Finetuning	26.89	24.80	27.38	–	–	–	–	–	–
ER	–	–	–	31.31	27.36	31.85	38.23	29.28	39.30
SAMM	–	–	–	36.27	24.98	35.74	42.53	28.12	43.72
AV-DreamNet *(Ours)*									
JOINT	88.29±0.33	80.76±0.49	92.14±0.69	–	–	–	–	–	–
Finetuning	83.56±0.22	77.04±0.25	87.82±0.36	–	–	–	–	–	–
Unimodal	86.39±0.22	79.26±0.19	–	85.61±0.34	77.54±0.18	–	85.86±0.33	78.12±0.62	–
Concat	–	–	90.84±0.13	–	–	90.49±0.44	–	–	**91.11±0.32**
Avg	–	–	**91.10±0.20**	–	–	**90.57±0.43**	–	–	91.04±0.31
FiLM	–	–	90.41±0.47	–	–	90.24±0.51	–	–	90.51±0.37

TABLE III: Results on GCIL-VGGSound (Generalized Class-Incremental Learning). A = Audio, V = Visual, AV = Audio-Visual accuracy.

Method	Buffer = 0			Buffer = 5			Buffer = 20		
	A	V	AV	A	V	AV	A	V	AV
SAMM [3]									
JOINT	44.02	26.23	49.32	–	–	–	–	–	–
Finetuning	20.34	11.47	24.73	–	–	–	–	–	–
ER	–	–	–	24.57	13.80	29.76	31.30	17.25	37.77
SAMM	–	–	–	27.41	11.90	34.34	31.96	14.35	42.08
AV-DreamNet *(Ours)*									
JOINT	58.76±0.41	47.86±0.46	67.50±0.45	–	–	–	–	–	–
Finetuning	15.63±0.65	7.30±0.27	16.39±0.52	–	–	–	–	–	–
Unimodal	42.05±0.75	25.81±3.20	–	45.52±0.55	40.06±0.51	–	42.60±0.81	37.49±0.58	–
Concat	–	–	**43.51±4.58**	–	–	**56.46±1.11**	–	–	**54.62±0.61**
Avg	–	–	43.38±1.32	–	–	54.38±0.75	–	–	51.80±0.39
FiLM	–	–	37.23±1.14	–	–	52.24±0.69	–	–	48.23±1.25

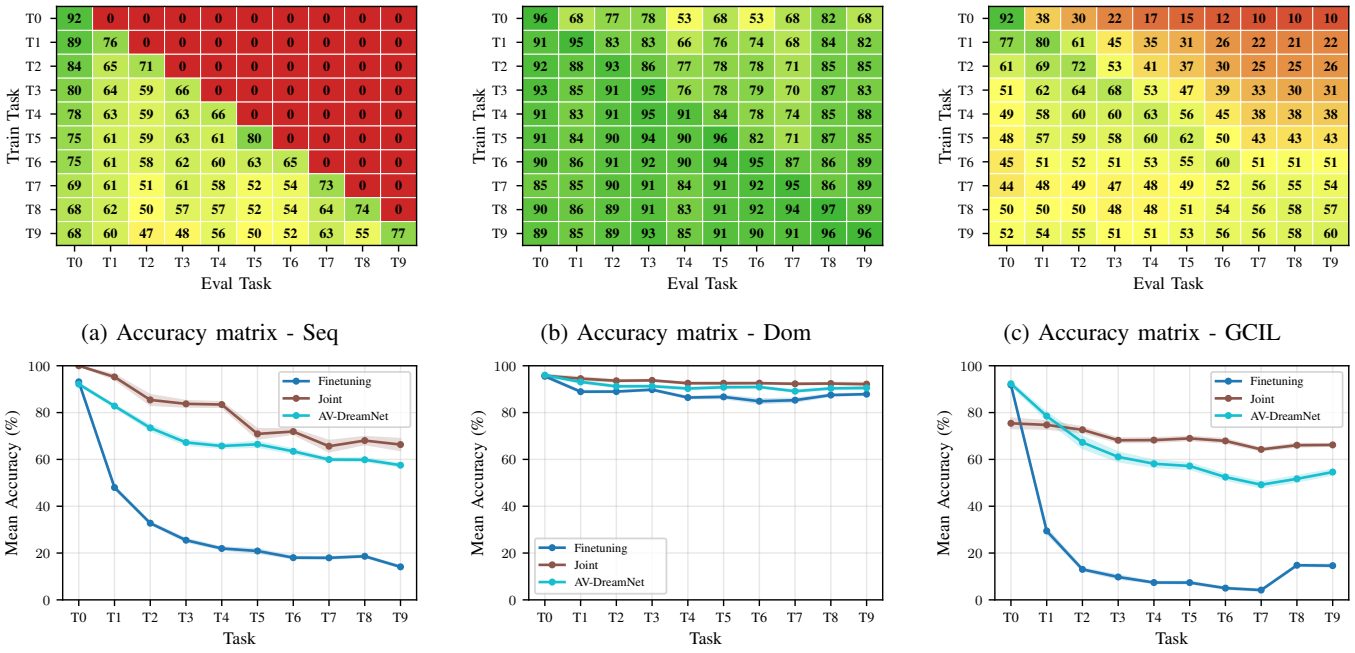
from PANNs for audio features (6.3M parameters including spectrograms extraction). Both encoders are frozen during learning, reducing trainable parameters to the MLP heads only (11M parameters for XS configuration). Total model size is approximately 29M parameters, a 5× parameter reduction compared to CLIP (150M parameters, vision-only).

4) Training and hyperparameter optimization: The learnable components consist of modality-specific encoder-decoder MLPs with average fusion for classification. We train using Adam optimizer for 3 epochs per task with batch size 64. For each experimental configuration including baselines, we perform hyperparameter optimization using the Tree-structured Parzen Estimator (TPE) algorithm [23] over 100 trials, optimizing the final accuracy score and systematically tuning learning rate, replay ratio r (limited to 6), and number of reinjections K (limited to 6). All final hyperparameter set are trained then evaluated on the test set over 10 runs with different random seeds, and we report mean ± standard deviation.

B. Main Results

1) Analysis: Tables I–III present results across the three scenarios. Our AV-DreamNet achieves substantially higher joint upper bounds than SAMM (69.7% vs 58.2% on Seq-VGGSound), indicating stronger multimodal representations from our frozen encoder setup.

On Seq-VGGSound (Table I), AV-DreamNet with buffer=5 achieves 58.1% accuracy, outperforming SAMM (26.3%) and ER (21.4%) by large margins. Even without any buffer (pure pseudo-rehearsal), our method reaches 47.1%, demonstrating



(a) Accuracy matrix - Seq (b) Accuracy matrix - Dom (c) Accuracy matrix - GCIL

(d) Mean accuracy - Seq (e) Mean accuracy - Dom (f) Mean accuracy - GCIL

Fig. 2: Accuracy matrices (top) and mean accuracy curves (bottom) across CL-VGGSound scenarios (buffer=5, average fusion). In matrices, entry (i, j) shows accuracy on task j after training on task i; below-diagonal entries reveal forgetting. Dom-VGGSound exhibits strong forward transfer due to shared supercategory structure. Curves show AV-DreamNet nearly compensates catastrophic forgetting, maintaining accuracy close to the Joint upper bound.

effective knowledge preservation through noise-based replay alone.

Dom-VGGSound (Table II) shows the strongest results, with AV-DreamNet achieving over 90% accuracy across all buffer sizes. This suggests that domain-incremental learning benefits particularly from pseudo-rehearsal, as the shared supercategory structure provides stable attractors for reinjection-based refinement.

On the challenging GCIL scenario (Table III), our method achieves 56.5% with buffer=5 compared to SAMM's 34.3%, maintaining competitive performance despite imbalanced task sizes. The accuracy matrices in Figure 2 reveal that AV-DreamNet maintains consistent per-task performance across the learning sequence, with limited forgetting of earlier tasks. The mean accuracy curves show that our method maintains a stable trajectory between the Joint upper bound and Finetuning lower bound throughout the task sequence.

C. Ablation Studies

1) Impact of Fusion Strategies: We compare three fusion mechanisms: concatenation, average, and FiLM. As shown in Tables I–III, concatenation and average fusion consistently outperform FiLM across all scenarios. On Seq-VGGSound with buffer=5, average fusion achieves 58.1% while FiLM reaches only 52.1%. This contrasts with SAMM's findings where FiLM excelled, suggesting that pseudo-rehearsal benefits from simpler fusion mechanisms that preserve modality-specific information for auto-associative reconstruction. The learnable parameters in FiLM may introduce instability when trained on noise-based pseudo-samples.

2) Impact of model depth: We evaluate four encoder-decoder architectures of increasing complexity:

- **XS**: 1×1024 encoders, 11,644,516 parameters and 11,661,712 FLOPs
- **S**: 1×1024 encoders + 512 classifier head, 12,118,116 parameters and 12,134,800 FLOPs
- **M**: 2×1024 encoders + 1×1024 decoders, 15,842,916 parameters and 15,856,016 FLOPs
- **L**: 2×1024 encoders + 1×1024 decoders + 512 classifier head, 16,316,516 parameters and 16,329,104 FLOPs

Figure 3 shows that shallower architectures (XS, S) consistently outperform deeper ones across all scenarios. This suggests that deeper networks have reduced capacity to reproduce the transfer function from noisy examples during pseudo-rehearsal. The XS configuration achieves the best trade-off, supporting our design choice of lightweight MLPs for embedded deployment.

3) Impact of Reinjection Steps and Replay Ratio: Figure 4 shows the interaction between reinjection steps K and replay ratio r on Seq-VGGSound. Both parameters independently improve accuracy: increasing K refines pseudo-samples toward stable attractors, while higher r provides more diverse replay examples. The total number of pseudo-samples per real example is $r \times K$, ranging from 1 (at $r = 1, K = 1$) to 256 (at

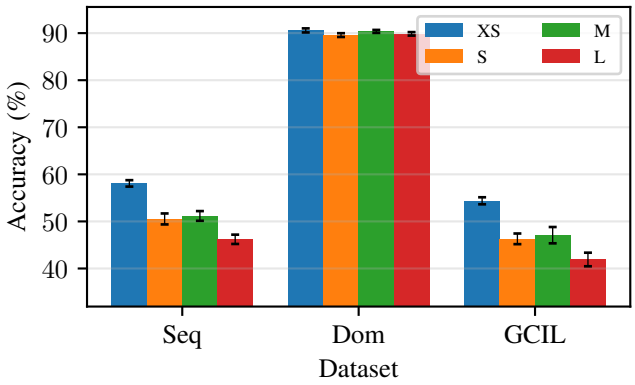


Fig. 3: Impact of model depth on accuracy. Shallow networks (XS) consistently outperform deeper configurations, suggesting deeper networks struggle to reproduce the transfer function from noisy pseudo-samples.

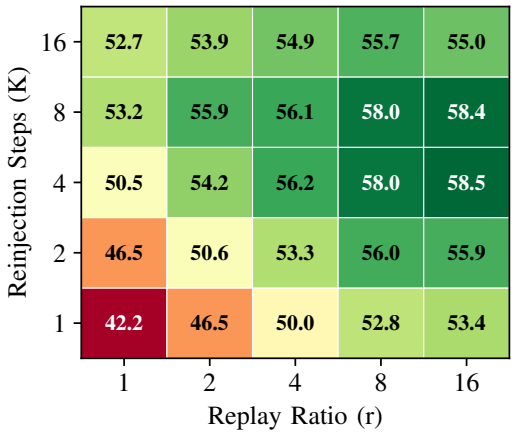


Fig. 4: Impact of reinjection steps and replay ratio on Seq-VGGSound accuracy. Both parameters improve pseudo-sample quality, with accuracy ranging from 42.2% (1 reinjection, ratio 1) to 58.5% (4 reinjections, ratio 16). Higher values yield better performance.

$r = 16, K = 16$). Accuracy improves from 42.2% to 58.5% across this range, with diminishing returns at higher values. For practical deployment, $K = 4$ and $r = 4$ provide a good trade-off, achieving 56.2% accuracy with 16 pseudo-samples per real example.

4) Impact of Buffer size: Figure 5 reveals a key finding: accuracy improves sharply with minimal buffers (1–5 samples per class) then plateaus, demonstrating that pseudo-rehearsal substantially reduces buffer requirements. On Seq-VGGSound, accuracy jumps from 47% (buffer=0) to 57% (buffer=5), but only marginally improves to 58% at buffer=40. This confirms the buffer-seeded replay hypothesis: a few real samples provide high-quality seeds for pseudo-generation, eliminating the need for large replay buffers. For Dom-VGGSound, performance is remarkably stable across buffer sizes, suggesting

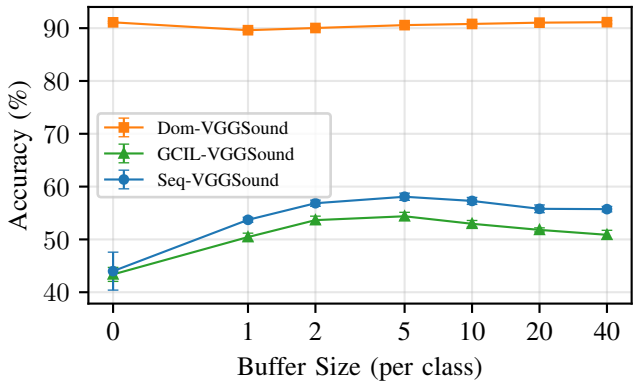


Fig. 5: Impact of buffer size on accuracy across the three CL scenarios. Accuracy improves with small buffers (1–5 samples per class) then plateaus, demonstrating that pseudo-rehearsal reduces the need for large replay buffers.

that domain-incremental scenarios benefit most from pseudo-rehearsal's ability to capture category-level structure.

V. Discussion

A. Privacy-Preserving Multimodal Learning

Our noise-based pseudo-rehearsal enables continual learning without storing sensitive audio-visual data, addressing GDPR and uses-cases requiring privacy. Unlike buffer-based methods, our approach generates synthetic representations that cannot be inverted to recover original data. Even in buffer-seeded mode, storing only a few latent samples per class when shuffling modalities significantly reduces privacy exposure, making AV-DreamNet suitable for privacy-critical uses-cases such as smart home devices, healthcare monitoring, and personal assistants.

B. Embedded Deployment Considerations

Our architecture is designed for resource-constrained deployment through several design choices. With approximately 29M total parameters, AV-DreamNet fits within the memory constraints of edge devices. The frozen encoder strategy is particularly advantageous for embedded systems: feature extraction can be offloaded to dedicated inference accelerators such as NeuroCorgi [24], which embeds fixed neural network weights directly in silicon to achieve ultra-low power consumption (36μJ/frame for ImageNet feature extraction). By keeping encoders frozen, continual learning occurs only in the lightweight MLP heads, enabling on-device adaptation without retraining the computationally expensive feature extractors.

Unlike VAE-based approaches [7], our noise-based pseudo-rehearsal requires no auxiliary generative model training. By generating pseudo-examples directly in latent space behind the frozen encoders, we avoid the need for heavy decoders to reconstruct samples in the original input space, greatly reducing the computational cost of example generation and reinjections during embedded continual learning.

Storing the buffer behind the frozen encoders in latent space drastically reduces th buffers memory footprint: ~600k values

per raw visual sample ($4 \times 224 \times 224 \times 3$) reduced to 2048 encoded features, and $\sim$320k values per raw audio sample (10×32k) reduced to 512 encoded features, yielding a $\sim 360\times$ overall reduction.

However, we store a frozen copy of the model head from the previous task to generate pseudo-examples, which doubles the model memory footprint during training. This remains significantly lighter than maintaining separate encoder-decoder generative architectures, but may require model compression techniques for the most constrained devices.

C. Limitations

The method's effectiveness relies on the frozen pretrained encoders, which may limit generalization to audio-visual domains significantly different from the pretraining data (ImageNet, AudioSet). Our evaluation is limited to CL-VGGSound; generalization to other audio-visual domains requires further validation. The gap to joint training upper bounds indicates room for improvement, and real-world embedded deployment remains future work.

VI. Conclusion

We presented AV-DreamNet, the first noise-based pseudo-rehearsal method for audio-visual continual learning. On CL-VGGSound, our approach outperforms buffer-based methods (58.1% vs 26.3% for SAMM on Seq-VGGSound with buffer=5) while enabling privacy-preserving learning without storing raw data. With lightweight frozen encoders enabling hardware acceleration and significant memory reduction through latent-space generation and buffers, AV-DreamNet provides a practical solution for embedded deployment. Future work includes extending to additional modalities, exploring others fusion methodes and deployment on actual embedded hardware.

Acknowledgment

The authors used Claude Opus 4.5 [25] as a writing assistant to improve vocabulary, sentence clarity, and paragraph conciseness. All scientific content, experimental design, and conclusions remain the sole responsibility of the authors.

References

[1] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *Psychology of Learning and Motivation*, vol. 24, pp. 109–165, 1989.

[2] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 114, no. 13, pp. 3521–3526, 2017.

[3] F. Sarfraz, B. Zonooz, and E. Arani, "Beyond unimodal learning: The importance of integrating multiple modalities for lifelong learning," in *Conference on Lifelong Learning Agents (CoLLAs)*, 2024, pp. 1–24.

[4] M. Mainsant, M. Solinas, M. Reyboz, C. Godin, and M. Mermillod, "Dream net: A privacy preserving continual learning model for face emotion recognition," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2021, pp. 1–8. [Online]. Available: https://hal.science/cea-03474722

[5] S. Mo, W. Pian, and Y. Tian, "Class-incremental grouping network for continual audio-visual learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 8868–8878.

[6] W. Pian, Y. Nan, S. Deng, S. Mo, Y. Guo, and Y. Tian, "Continual audio-visual sound separation," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024.

[7] E. Frascaroli, A. Panariello, P. Buzzega, L. Bonicelli, A. Porrello, and S. Calderara, "CLIP with generative latent replay: a strong baseline for incremental learning," in *British Machine Vision Conference (BMVC)*, 2024.

[8] M. Solinas, M. Reyboz, and M. Mermillod, "Beneficial effect of combined replay for continual learning," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART)*, vol. 2, 2021, pp. 733–740.

[9] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *International Conference on Machine Learning (ICML)*. PMLR, 2017, pp. 3987–3995.

[10] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 40, no. 12, pp. 2935–2947, 2017.

[11] S. Yan, J. Xie, and X. He, "DER: Dynamically expandable representation for class incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3014–3023.

[12] F.-Y. Wang, D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "FOSTER: Feature boosting and compression for class-incremental learning," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 398–414.

[13] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2001–2010.

[14] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, "Dark experience for general continual learning: a strong, simple baseline," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 15 920–15 930.

[15] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 2990–2999.

[16] G. M. van de Ven, H. T. Siegelmann, and A. S. Tolias, "Brain-inspired replay for continual learning with artificial neural networks," *Nature Communications*, vol. 11, no. 1, p. 4069, 2020.

[17] R. Gao, X. Yu, H. Wang, M. Gao, M. Tong, X. Wang, H. Peng, and F. Li, "DDGR: Continual learning with deep diffusion-based generative replay," in *International Conference on Machine Learning (ICML)*, 2023, pp. 10 758–10 779.

[18] L. Pellegrini, D. Trappetti, M. Masone, G. Bendotti, A. Carta, D. Bacciu, and D. Maltoni, "Latent generative replay for resource-efficient continual learning of facial expressions," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, 2023, pp. 1–8.

[19] M. Mainsant, M. Solinas, M. Reyboz, and M. Mermillod, "Robustness study of continual learning approach for facial expression recognition model on fer-2013 with different incremental learning scenarios," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2022, pp. 1093–1098. [Online]. Available: https://hal.science/cea-03982204

[20] M. Solinas, M. Reyboz, and M. Mermillod, "On the beneficial effects of reinjections for continual learning," *SN Computer Science*, vol. 4, no. 3, p. 263, 2023. [Online]. Available: https://hal.science/hal-03924062

[21] X. Yue, X. Zhang, Y. Chen, C. Zhang, M. Lao, H. Zhuang, X. Qian, and H. Li, "MMAL: Multi-modal analytic learning for exemplar-free audio-visual class incremental tasks," in *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, 2024, pp. 4959–4968.

[22] X. He, D. Zhao, Y. Li, Q. Kong, X. Yang, and Y. Zeng, "Incorporating brain-inspired mechanisms for multimodal learning in artificial intelligence," *arXiv preprint arXiv:2505.10176*, 2025.

[23] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 24, 2011, pp. 2546–2554.

[24] I. Miró-Panadès, V. Lorrain, L. Billod, I. Kucher, V. Templier, S. Choisnet, N. Ali, B. Rossigneux, O. Bichler, and A. Valentian, "A 7072j/frame ImageNet feature extractor accelerator on HD images at 30FPS," in *IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*. IEEE, 2024.

[25] Anthropic, "Claude opus 4.5," 2025, large language model. [Online]. Available: https://www.anthropic.com/claude