EWC-Guided Diffusion Replay for Exemplar Free Continual Learning in Medical Imaging

Anoushka Harit[1]
ah2415@cam.ac.uk

William Prew[1]
wp287@cam.ac.uk

Zhongtian Sun[1,2]
z.sun-256@kent.ac.uk

Florian Markowetz[1]
fm327@cam.ac.uk

[1]**University of Cambridge, Cambridge, United Kingdom**
[2]**University of Kent, Canterbury, United Kingdom**

Abstract

Medical imaging foundation models must adapt over time, yet full retraining is often blocked by privacy constraints and cost. We present a continual learning framework that avoids storing patient exemplars by pairing class conditional diffusion replay with Elastic Weight Consolidation. Using a compact Vision Transformer backbone, we evaluate across eight MedMNIST v2 tasks and CheXpert. On CheXpert our approach attains 0.851 AUROC, reduces forgetting by more than 30% relative to DER++, and approaches joint training at 0.869 AUROC, while remaining efficient and privacy preserving. Analyses connect forgetting to two measurable factors: fidelity of replay and Fisher weighted parameter drift, highlighting the complementary roles of replay diffusion and synaptic stability. The results indicate a practical route for scalable, privacy aware continual adaptation of clinical imaging models.

1 Introduction

Foundation models (FMs) are reshaping medical AI by enabling transferable representations across imaging tasks Wu et al. [2025]. However, in real-world deployment, they must continually adapt to new diseases, imaging protocols, and workflows. Retraining is costly, and storing patient data for replay is often restricted, making privacy-preserving continual learning essential. This need intensifies *catastrophic forgetting* McCloskey and Cohen [1989], where learning new tasks erodes prior knowledge. Existing solutions fall short: regularisation (e.g., EWC Kirkpatrick et al. [2017], EFT Liu et al. [2022]) degrades under distribution shift; exemplar-based replay (e.g., DER++ Buzzega et al. [2020], SPM Zhu et al. [2021]) violates privacy; generative replay (e.g., VAEs, GANs Shin et al. [2017], Van de Ven et al. [2020]) struggles with fine medical detail; and dynamic expansion (e.g., PMoE Jung and Kim [2024], CoPE De Lange and Tuytelaars [2021]) incurs high compute. **Our approach.** We propose a hybrid framework that combines a lightweight Vision Transformer (ViT), class-conditional DDPM for exemplar-free replay, and EWC for stability. Diffusion preserves subtle medical detail without storing data, and the unified model handles both 2D and 3D tasks efficiently. We also derive a forgetting bound linking replay fidelity and parameter anchoring. Experiments on MedMNIST v2 Yang et al. [2023] and CheXpert Irvin et al. [2019] confirm strong retention across modalities Sun et al. [2022].

Contributions. (i) A privacy-preserving continual learning framework combining DDPM replay and EWC; (ii) unified handling of 2D/3D medical tasks with a single diffusion model; (iii) a theoretical

bound linking forgetting, replay fidelity, and parameter stability; (iv) empirical validation on multi-scale benchmarks.

2 Continual Learning Setup

Let $\mathcal{T} = \{T_1, \ldots, T_K\}$ be a sequence of tasks with datasets $D_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{n_k}$, where $x_i^{(k)} \in \mathcal{X}$ and $y_i^{(k)} \in \mathcal{Y}_k$. At step k, only D_k is available; D_j for $j < k$ cannot be stored (exemplar-free constraint).

A model $f_\theta : \mathcal{X} \to \Delta(\mathcal{Y})$ with parameters $\theta \in \mathbb{R}^d$ is updated sequentially $\theta^{(k)} \mapsto \theta^{(k+1)}$.

Forgetting is measured as

$$\mathcal{F} = \tfrac{1}{K-1} \sum_{j=1}^{K-1} \max_{t \leq K} \left(A_{j,t} - A_{j,K} \right),$$

and final accuracy as

$$\mathcal{A} = \tfrac{1}{K} \sum_{j=1}^{K} A_{j,K},$$

where $A_{j,t}$ is accuracy on T_j after step t.

The objective is to maximise $\mathcal{A}$ while minimising $\mathcal{F}$ under exemplar-free continual learning.

3 Theoretical Analysis of Forgetting

We analyse forgetting in exemplar-free continual learning by decomposing it into two measurable sources that directly motivate our method: (i) *distributional drift*, arising from imperfect replay, and (ii) *parameter drift*, arising from unstable optimisation.
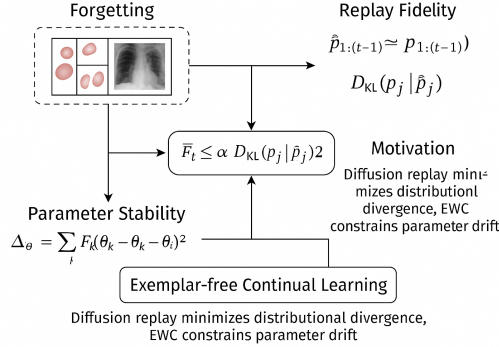

Figure 1: Problem formulation of exemplar-free continual learning. Past datasets $\mathcal{D}_{1:(t-1)}$ cannot be stored; the learner sees only the current task $\mathcal{D}_t$ and must perform well across all $\mathcal{D}_{1:T}$. Two drivers of forgetting emerge: replay fidelity (divergence $D_{\text{KL}}(p_j \| \hat{p}_j)$) and parameter stability (Fisher-weighted drift $\sum_i F_i(\theta_i - \theta_i^\star)^2$).

Setup. For each task j, let $\mathcal{A}_j^\star$ denote accuracy immediately after training, and $\mathcal{A}_j^T$ the accuracy after learning all T tasks. Forgetting on task j is

$$F_j = \mathcal{A}_j^\star - \mathcal{A}_j^T, \qquad \bar{F} = \tfrac{1}{T} \sum_{j=1}^{T} F_j.$$

Distributional drift. When replay substitutes a proxy $\hat{p}_j$ for the true p_j, Pinsker's inequality for bounded loss $\ell \in [0, L_{\max}]$ gives

$$\left| \mathbb{E}_{p_j}[\ell] - \mathbb{E}_{\hat{p}_j}[\ell] \right| \leq L_{\max} \sqrt{\tfrac{1}{2} D_{\text{KL}}(p_j \| \hat{p}_j)}.$$

2

Thus, replay error contributes in proportion to the KL divergence. This term motivates *diffusion replay*, which yields lower divergence than VAEs or GANs, and *Fisher Scheduled Replay*, which allocates generative samples where divergence is most damaging.

Parameter drift. Let $\theta_j^\star$ be the optimum for task j. A second-order expansion around $\theta_j^\star$ gives

$$\mathcal{L}(\theta) \approx \mathcal{L}(\theta_j^\star) + \tfrac{1}{2}(\theta - \theta_j^\star)^\top F(\theta - \theta_j^\star),$$

where F is the Fisher information matrix. The excess loss scales with

$$D_j = \sum_i F_i(\theta_i - \theta_i^\star)^2,$$

capturing instability of Fisher-salient parameters. This motivates *Elastic Weight Consolidation*, which explicitly penalises this Fisher-weighted drift.

Unified bound. Combining the two effects, forgetting can be bounded as

$$\bar{F} \;\leq\; \alpha \, D_{\mathrm{KL}}(p_j \,\|\, \hat{p}_j) + \beta \sum_i F_i(\theta_i - \theta_i^\star)^2,$$

with constants $\alpha, \beta > 0$ depending on loss smoothness and curvature. This bound maps directly to our design: diffusion replay reduces the KL term, FSR further focuses replay where divergence is largest, and EWC constrains the Fisher-weighted drift.

Empirical validation. Although α and β are not directly observable, both terms of the bound can be estimated. For each task j we compute replay divergence $\widehat{D}_{\mathrm{KL}}(p_j \,\|\, \hat{p}_j)$ and Fisher-weighted drift D_j, and relate them to observed forgetting via

$$F_j = a \, \widehat{D}_{\mathrm{KL}}(p_j \,\|\, \hat{p}_j) + b \, D_j + \varepsilon_j.$$

As reported in Appendix B.3, both terms correlate positively with forgetting, and the joint regression explains more variance than either alone. This provides empirical support for the replay–drift decomposition and justifies the integrated design of EWC-DR.

4   Methodology

We propose a **hybrid continual learning framework** for *privacy-preserving, exemplar-free* adaptation of medical imaging models. It integrates class-conditional DDPM replay to synthesise prior-task data, EWC to preserve critical parameters, and a lightweight ViT backbone to fuse real and replayed samples. Inspired by dual-memory theory McClelland et al. [1995], DDPM enables fast recall, while EWC supports gradual consolidation.
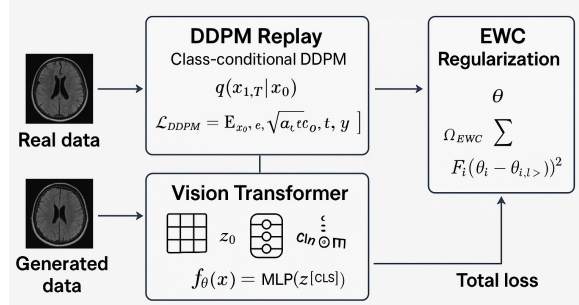

Figure 2: Hybrid continual learning: DDPM replay generates exemplar-free samples, fused with real data in a ViT classifier, with EWC for parameter stability.

4.1   Problem Setting

Let tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_K\}$ have datasets $\mathcal{D}_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{N_k}$ from evolving, unknown $\mathcal{P}_k(x, y)$. At step k, only $\mathcal{D}_k$ is accessible. The goal is to learn f_θ that generalises to $\mathcal{P}_{\leq k} = \bigcup_{j=1}^k \mathcal{P}_j$ without storing raw past data.

4.2 Overall Objective

We minimise:

$$\mathcal{L}_{\text{total}}^{(k)} = \mathbb{E}_{(x,y)\sim\mathcal{D}_k\cup\hat{\mathcal{D}}_{<k}}[\mathcal{L}_{\text{CE}}(f_\theta(x),y)] + \lambda\sum_i F_i(\theta_i - \theta_{i,<k}^*)^2, \tag{1}$$

where $\hat{\mathcal{D}}_{<k}$ are DDPM-generated samples, F_i are Fisher information entries, and $\theta_{i,<k}^*$ are anchored parameters from earlier tasks.

Component 1: DDPM Replay For each task $\mathcal{T}_k$, we train a single class-conditional DDPM $p_k(x \mid y)$ to reverse the diffusion process:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^T \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{2}$$

with training objective:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{x_0,y,\epsilon,t}\big\|\epsilon - \epsilon_\phi\big(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, y\big)\big\|^2. \tag{3}$$

The generated set $\hat{\mathcal{D}}_k$ is added to the replay buffer for future tasks.

Component 2: ViT Classifier An image $x \in \mathbb{R}^{H\times W\times C}$ is split into N patches $x_p^{(i)} \in \mathbb{R}^{P^2 C}$:

$$z_0 = [x_p^{(1)}\mathbf{E}; \dots; x_p^{(N)}\mathbf{E}] + \mathbf{E}_{\text{pos}}, \tag{4}$$

processed through L transformer layers, with [CLS] token $z_L^{[\text{CLS}]}$ classified via:

$$f_\theta(x) = \text{MLP}(z_L^{[\text{CLS}]}). \tag{5}$$

Component 3: EWC Regularisation EWC penalises drift on important parameters:

$$\Omega_{\text{EWC}} = \sum_i F_i(\theta_i - \theta_i^*)^2, \tag{6}$$

with F_i estimated on past tasks.

Class-Incremental Learning (CIL) CIL requires prediction over $\mathcal{C}_{\leq k}$ without task IDs. Our replay buffer maintains class balance, mitigating bias. Eq. (1) is applied with cross-entropy over the cumulative label set.

Theoretical View: Forgetting Bound We bound average forgetting as:

$$\bar{F} \leq \alpha\,\text{KL}(\mathcal{P}_k \parallel \hat{\mathcal{P}}_k) + \beta\,\|\theta_k - \theta_{k-1}^*\|_2^2, \tag{7}$$

linking replay fidelity (first term) and parameter stability (second term), motivating joint DDPM replay and EWC.

5 Experimental Setup and Results

Datasets. We evaluate on three benchmarks spanning distinct modalities: **MedMNIST-2D** Yang et al. [2023] [1] (six 2D tasks: PathMNIST, BloodMNIST, DermaMNIST, RetinaMNIST, BreastMNIST, PneumoniaMNIST), **MedMNIST-3D** (OrganMNIST3D, NoduleMNIST3D), and **CheXpert** Irvin et al. [2019][2] (14-label chest X-ray classification). A class-incremental protocol is used, with disjoint label sets per task and no task identity at test time.

---

[1]https://medmnist.com/

[2]https://aimi.stanford.edu/datasets/chexpert-chest-x-rays

Baselines. We compare against regularisation-based methods (*EFT* Liu et al. [2022], *CoPE* De Lange and Tuytelaars [2021]) that avoid replay but struggle under distribution shift; replay-based methods (*DER++* Buzzega et al. [2020], *SPM* Zhu et al. [2021], *VAE+Replay* Shin et al. [2017]); and the dynamic-expansion method *PMoE* Jung and Kim [2024]. All baselines use the same ViT backbone and training setup.

Implementation. We use a lightweight ViT (patch size 16, 6 layers, 8 heads, hidden dim 512) initialised from ImageNet. The DDPM is class-conditional (T=1000, cosine β_t) and trained for 200 epochs per task. EWC uses Fisher information from 500 samples per class with $\lambda \in \{10, 50, 100\}$ tuned on validation. Replay generates 256 class-balanced samples per task, stored as raw float32 tensors within a fixed 100 MB budget (Appendix H). Each batch mixes real and replayed data 1:1. All models use AdamW (lr 3×10^{-4}, weight decay 0.01), batch size 64, and results are averaged over five seeds.

Metrics. We report *Average Accuracy* after the final task, *Average Forgetting* Chaudhry et al. [2018], and *AUC* to reflect clinical relevance in multi-label settings. On CheXpert, we report mean AUC averaged across the 14 labels (macro-average).

Protocols. All methods are trained under a fixed canonical task order. Forgetting is measured from intermediate evaluations after each task and reported at the final step. For CheXpert, we report the mean AUC, macro-averaged across the 14 disease labels. Robustness to alternative task orders is analysed separately in Appendix G.

5.1 Main Results

Table 1 reports performance across MedMNIST-2D, MedMNIST-3D, and CheXpert. All methods use the same ViT backbone, memory budget (100 MB), and task sequence for fair comparison. **DDPM+EWC** consistently achieves the best trade-off between accuracy, AUC, and forgetting, outperforming regularisation-only methods (EWC, CoPE) and replay-based baselines (VAE+Replay, DER++). On CheXpert, it reduces forgetting by over 30% compared to DER++, underscoring the value of high-fidelity generative replay in clinically sensitive settings. Joint training remains an upper bound but violates continual learning constraints.[3]

Table 1: **Continual Adaptation Results.** Mean over 5 runs. Best in **bold**, second-best underlined. Forgetting ($F \downarrow$) lower is better.

Method	MedMNIST-2D			MedMNIST-3D			CheXpert		
	Acc↑	F↓	AUC↑	Acc↑	F↓	AUC↑	Acc↑	F↓	AUC↑
Finetune	67.4	27.5	0.820	63.2	29.1	0.801	64.8	26.9	0.802
EWC	72.9	19.7	0.842	68.5	21.5	0.824	70.5	19.4	0.824
EFT	71.1	21.4	0.839	67.9	22.9	0.820	69.4	20.5	0.820
CoPE	72.4	19.9	0.843	68.8	21.3	0.826	70.8	19.2	0.826
DER++	<u>75.6</u>	<u>14.2</u>	<u>0.853</u>	70.9	<u>16.8</u>	0.835	<u>73.2</u>	<u>13.8</u>	<u>0.838</u>
SPM	74.9	15.0	0.852	<u>71.2</u>	17.0	<u>0.836</u>	72.6	14.4	0.835
VAE+Replay	74.2	15.6	0.851	70.8	17.5	0.837	71.7	15.1	0.833
PMoE	74.5	15.3	0.852	70.9	17.2	0.835	72.1	14.7	0.834
Ours (DDPM+EWC)	**78.1**	**10.5**	**0.866**	**74.0**	**12.9**	**0.849**	**76.4**	**10.9**	**0.851**
Joint (Upper Bound)	81.4	0.0	0.879	77.5	0.0	0.861	79.1	0.0	0.869

5.2 Task-wise Performance

We report accuracy on the first (T_1), middle (T_{mid}), and final (T_n) tasks after completing the sequence. For MedMNIST-2D, T_1 and T_2 denote datasets (e.g., BreastMNIST, PneumoniaMNIST); for MedMNIST-3D, they refer to OrganMNIST3D and NoduleMNIST3D; and for CheXpert, they correspond to early, median, and final label groups. **DDPM+EWC** consistently preserves T_1 performance, particularly in CheXpert where strong domain shift typically causes rapid forgetting (Table 2).

---

[3]Finetune denotes sequential training without replay or regularisation and serves as a lower-bound baseline.

Table 2: **Task-wise accuracy (%)** after the final task for MedMNIST-2D, MedMNIST-3D, and CheXpert. Best in **bold**, second-best underlined.

Method	MedMNIST-2D			MedMNIST-3D			CheXpert		
	T_1	T_{mid}	T_n	T_1	T_{mid}	T_n	T_1	T_{mid}	T_n
Finetune	45.2	60.1	83.4	41.7	56.5	79.0	43.8	58.0	80.2
EWC	57.5	66.3	84.1	53.2	62.9	80.5	54.9	64.2	81.0
EFT	55.9	65.4	83.8	52.4	61.7	80.2	53.6	63.2	80.6
CoPE	58.4	67.1	84.0	54.5	63.8	80.6	56.0	64.9	81.2
DER++	<u>63.9</u>	<u>70.8</u>	84.5	<u>59.0</u>	<u>67.8</u>	<u>81.2</u>	<u>61.5</u>	<u>69.1</u>	<u>82.3</u>
SPM	62.7	70.1	84.3	58.4	67.1	81.0	60.2	68.7	82.0
VAE+Replay	61.5	69.7	84.2	57.8	66.5	80.8	59.8	68.0	81.7
PMoE	62.0	69.9	<u>84.4</u>	58.1	66.9	81.1	60.5	68.2	81.9
Ours (DDPM+EWC)	**67.8**	**73.2**	**85.1**	**62.3**	**70.4**	**82.5**	**65.7**	**72.4**	**83.4**

5.3 Ablation Study

Table 3 shows the impact of removing DDPM or EWC. Excluding DDPM increases forgetting by 5.8% (2D), 6.6% (3D), and 7.4% (CheXpert), highlighting the role of replay fidelity. Removing EWC reduces early-task accuracy by up to 4.9%, 4.5%, and 3.7%, confirming the importance of parameter stability.

Table 3: **Ablation results** on MedMNIST-2D, MedMNIST-3D, and CheXpert.

Method Variant	MedMNIST-2D		MedMNIST-3D		CheXpert	
	Avg Acc ↑	Forget ↓	Avg Acc ↑	Forget ↓	Avg Acc ↑	Forget ↓
Full Model (DDPM+EWC)	**72.8**	**11.3**	**68.9**	**12.4**	**68.5**	**13.7**
w/o DDPM (EWC only)	67.0	17.1	63.5	18.0	62.3	21.1
w/o EWC (DDPM only)	69.2	14.5	65.4	15.7	64.8	18.9

6 Conclusion and Impact

We introduced an exemplar-free continual learning framework that couples diffusion-based replay with Elastic Weight Consolidation to retain prior knowledge without storing patient data. Across MedMNIST (2D and 3D) and CheXpert, the approach improves final accuracy, reduces forgetting against strong replay and regularisation baselines, and approaches joint training under a fixed memory budget. Our analysis links forgetting to replay divergence and Fisher-weighted parameter drift, providing actionable diagnostics for when and why degradation occurs. The method therefore offers a practical route to privacy-preserving adaptation of clinical imaging models.

This work advances trustworthy continual adaptation in medical imaging by (i) eliminating raw exemplar retention, (ii) providing quantitative signals for governance (calibration, drift traces, and replay quality), and (iii) operating within realistic memory limits. It enables hospitals to update models as distributions shift while respecting privacy constraints and audit requirements. Future directions include stronger calibration, fairness under class and site imbalance, compute-efficient replay via generator distillation, multi-modal extensions (e.g., image plus report), and prospective clinical validation.

References

Jonas Adler and Sebastian Lunz. Banach wasserstein gan. *Advances in Neural Information Processing Systems*, 31, 2018.

Rahaf Aljundi and et al. Memory aware synapses: Learning what (not) to forget. In *European Conference on Computer Vision*, 2018.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. URL https://arxiv.org/abs/1804.03599.

Pietro Buzzega, Matteo Boschini, Alessandro Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In *NeurIPS*, 2020.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547, 2018.

Matthias De Lange and Tinne Tuytelaars. Continual prototype evolution: Learning online from non-stationary data streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8250–8259, 2021.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Min Jae Jung and JooHee Kim. Pmoe: Progressive mixture of experts with asymmetric transformer for continual learning, 2024. URL https://arxiv.org/abs/2407.21571.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. In *Proceedings of the National Academy of Sciences*, volume 114, pages 3521–3526, 2017.

Xu Liu et al. Elastic feature transformation for continual learning. In *ICLR*, 2022.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419, 1995.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2990–2999, 2017.

Zhongtian Sun, Anoushka Harit, Alexandra I Cristea, Jialin Yu, Noura Al Moubayed, and Lei Shi. Is unimodal bias always bad for visual question answering? a medical domain study with dynamic attention. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5352–5360. IEEE, 2022.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

Gido M Van de Ven, Hava T Siegelmann, and Andreas S Tolias. Brain-inspired replay for continual learning with artificial neural networks. *Nature communications*, 11(1):4069, 2020.

RunQing Wu, KaiHui Huang, HanYi Zhang, QiHe Liu, JinYu Guo, JingSong Deng, and Fei Ye. Information-theoretic dual memory system for continual learning. *Knowledge-Based Systems*, 325: 113913, 2025.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017.

Yanzhao Zhu, Yuting Liu, Jianmin Liang, Tianqi Xu, and Jian Zhang. Prototype contrastive learning for efficient class-incremental learning. In *ICCV*, 2021.

Appendix / Supplemental Results

A Related Work

Continual learning (CL) addresses the problem of catastrophic forgetting, where neural networks lose previously acquired knowledge when trained on new tasks. This limitation presents significant challenges in high-stakes domains such as medical imaging, particularly for benchmarks such as MedMNIST-2D , MedMNIST-3D Yang et al. [2023], and CheXpert Irvin et al. [2019], where accurate retention of past diagnostic features is critical McCloskey and Cohen [1989], Parisi et al. [2019].

Regularisation-based methods aim to preserve prior knowledge by constraining parameter updates. Elastic Weight Consolidation (EWC) estimates the importance of network weights using the Fisher Information matrix and penalises deviations from their previous values during training on new tasks Kirkpatrick et al. [2017]. This framework, inspired by synaptic consolidation in biological systems McClelland et al. [1995], has been extended by methods such as Synaptic Intelligence Zenke et al. [2017] and Memory Aware Synapses Aljundi and et al. [2018], which perform online importance estimation. While regularisation methods are memory efficient, they may struggle in settings with substantial domain shift or class imbalance Chaudhry et al. [2018].

Replay-based approaches address forgetting by reintroducing prior knowledge, either by storing past examples or generating synthetic ones. Exemplar-based methods such as iCaRL store a small subset of prior data and interleave them with new inputs during training [Rebuffi et al., 2017]. Generative replay eliminates the need for data storage by synthesising samples from previous tasks. Deep Generative Replay (DGR) demonstrated this approach using generative adversarial networks (GANs) Shin et al. [2017], though GANs are known to suffer from mode collapse and unstable training Adler and Lunz [2018]. Variational autoencoders (VAEs) offer a more stable alternative but frequently generate blurry reconstructions due to pixel-wise loss, which is particularly limiting in medical imaging tasks that rely on fine-grained features Kingma et al. [2013], Burgess et al..

Diffusion models have recently emerged as a strong alternative to GANs and VAEs, offering superior stability and sample quality. Denoising diffusion probabilistic models (DDPMs) learn to reverse a gradual noising process and have achieved state-of-the-art results in image generation tasks Ho

et al. [2020], Dhariwal and Nichol [2021]. Their high fidelity and diversity make them promising for medical replay, where preserving subtle clinical patterns is essential. Diffusion models have shown success in applications such as segmentation and anomaly detection in radiology and histopathology Kazerouni et al. [2023], but, to our knowledge, there is no prior work applying DDPM replay to exemplar-free continual adaptation in medical imaging.

In addition, while convolutional networks have traditionally been the default backbone in continual learning, recent work has shown that Vision Transformers (ViTs) can offer improved flexibility and transferability across tasks. ViTs operate on sequences of image patches, making them naturally suited to capture both global and local patterns that vary across medical datasets Dosovitskiy et al. [2020]. In our work, we integrate a lightweight Vision Transformer as a domain-specific foundation model backbone within our continual learning framework. This design complements the high-fidelity generative replay of DDPMs and the synaptic stability provided by EWC, offering a robust and scalable architecture for continual adaptation of medical foundation models across diverse modalities Touvron et al. [2021].

B Theoretical Justifications and Bound

We provide a theoretical analysis of our hybrid continual adaptation framework. Our analysis links forgetting to two fundamental factors in FM adaptation: (i) **distributional shift** between original and replayed data, and (ii) **parameter drift** in the FM backbone. This aligns with the workshop themes of understanding generalisation forgetting trade-offs on evolving distributions and providing theoretical guarantees for continual FM updates.

B.1 Elastic Weight Consolidation as a Bayesian Prior

EWC can be interpreted as an online Laplace approximation of the posterior $p(\theta|\mathcal{D}_{1:k})$ over model parameters for the FM backbone. Let $\mathcal{L}_k(\theta) = -\log p(\mathcal{D}_k|\theta)$ denote the task loss. The EWC objective is:

$$\mathcal{L}_{\text{EWC}}(\theta) = \mathcal{L}_k(\theta) + \frac{\lambda}{2} \sum_i F_i(\theta_i - \theta_i^*)^2, \tag{8}$$

where F_i is the diagonal Fisher information and θ^* are parameters from prior adaptation stages. Expanding the log-prior around θ^* via a second-order Taylor expansion yields:

$$\log p(\mathcal{D}_{1:k-1}|\theta) \approx \log p(\mathcal{D}_{1:k-1}|\theta^*) - \frac{1}{2}(\theta - \theta^*)^\top F(\theta - \theta^*), \tag{9}$$

making explicit the role of EWC in anchoring parameters that are critical for past FM capabilities.

B.2 Distributional Stability of DDPM Replay

Let $\hat{\mathcal{D}}_j$ denote synthetic samples from a task-specific DDPM g_{ϕ_j} trained on $\mathcal{D}_j$. The FM classifier risk after k tasks is:

$$\mathcal{R}_k = \mathbb{E}_{(x,y) \sim \mathcal{D}_{1:k}}[\ell(f_\theta(x), y)]. \tag{10}$$

Assume:

- Total variation distance $D_{\text{TV}}(\mathcal{D}_j, \hat{\mathcal{D}}_j) \le \epsilon$,
- Loss ℓ is L-Lipschitz.

Then the risk gap between real and replayed data satisfies:

$$|\mathcal{R}_k - \hat{\mathcal{R}}_k| \le L \cdot \epsilon \cdot \sum_{j=1}^{k-1} \frac{|\hat{\mathcal{D}}_j|}{|\mathcal{D}_{1:k}|}. \tag{11}$$

This formalises the CCFM workshop theme that high-fidelity replay is crucial to mitigating forgetting in FMs continually updated on non-stationary data.

B.3 Derivation of the Forgetting Bound

We define the average forgetting across K adaptation steps as:

$$\bar{F} = \frac{1}{K} \sum_{k=1}^{K} \left[A_k - A_{k,K} \right], \tag{12}$$

where A_k is accuracy on task k immediately after learning and $A_{k,K}$ is accuracy after completing all K tasks. We decompose $\bar{F}$ into **distributional shift** and **parameter shift** terms.

Distributional Shift Term. If $\ell(x,y) \in [0, L_{\max}]$, then using Pinsker's inequality:

$$\left| \mathbb{E}_{\mathcal{D}_k}[\ell] - \mathbb{E}_{\hat{\mathcal{D}}_k}[\ell] \right| \leq L_{\max} \|\mathcal{D}_k - \hat{\mathcal{D}}_k\|_{\mathrm{TV}} \leq L_{\max} \sqrt{\tfrac{1}{2} \mathrm{KL}(\mathcal{D}_k \| \hat{\mathcal{D}}_k)}, \tag{13}$$

highlighting that greater KL divergence in replay worsens retention—central to the workshop's focus on temporal shift modelling.

Parameter Shift Term. Let $\theta_k^* = \arg\min_\theta \mathcal{L}(\theta)$ for $\mathcal{D}_k$. A second-order Taylor expansion and $H \approx F$ gives:

$$\mathcal{L}(\theta) \approx \mathcal{L}(\theta_k^*) + \tfrac{1}{2} \sum_i F_i (\theta_i - \theta_{k,i}^*)^2, \tag{14}$$

quantifying drift from parameters optimal for earlier FM tasks.

Combined Bound. Combining both terms yields:

$$A_k - A_{k,K} \lesssim \alpha \, \mathrm{KL}(\mathcal{D}_k \| \hat{\mathcal{D}}_k) + \beta \, \|\theta_K - \theta_k^*\|_2^2, \tag{15}$$

with $\alpha = L_{\max}/\sqrt{2}$ and $\beta = \frac{1}{2} \max_i F_i$. Averaging gives:

$$\bar{F} \lesssim \frac{1}{K} \sum_{k=1}^{K} \left[\alpha \, \mathrm{KL}(\mathcal{D}_k \| \hat{\mathcal{D}}_k) + \beta \, \|\theta_K - \theta_k^*\|_2^2 \right]. \tag{16}$$

This formalises the trade-off between generative replay fidelity and parameter stability for evolving FMs.

B.4 Analytical Bound under Smoothness Assumptions

Under $D_{\mathrm{KL}}(p\|\hat{p}) \leq \delta$ and large λ:

$$\mathbb{E}[F_k] \leq \alpha\delta + \beta\lambda^{-1}, \tag{17}$$

implying that $\delta \to 0$ (high-fidelity replay) and $\lambda \to \infty$ (strong parameter anchoring) jointly minimise forgetting, a central objective in CCFM's vision for reliable FM updates.

C Training Algorithm for Foundation Model Continual Adaptation

This section describes the complete training algorithm for our hybrid continual learning framework, which integrates Denoising Diffusion Probabilistic Models (DDPMs) for replay and Elastic Weight Consolidation (EWC) for synaptic regularisation. The design aims to mitigate catastrophic forgetting in medical foundation models while preserving scalability and patient data privacy. The method comprises five stages, detailed below and summarised in Algorithm 1.

1. **Task-specific diffusion model training:** For each new task, train a class-conditional DDPM g_{ϕ_k} to model $p(x|y)$ using the task dataset $\mathcal{D}_k$ without storing raw data.

2. **Synthetic sample generation:** Use the trained DDPM to generate replay samples $\hat{\mathcal{D}}_k$ from class labels y and add them to a generative replay buffer $\mathcal{M}$.

3. **Fisher Information Matrix estimation:** Compute a diagonal approximation of the empirical Fisher Information Matrix over $\mathcal{D}_k$ to quantify the importance of each parameter.
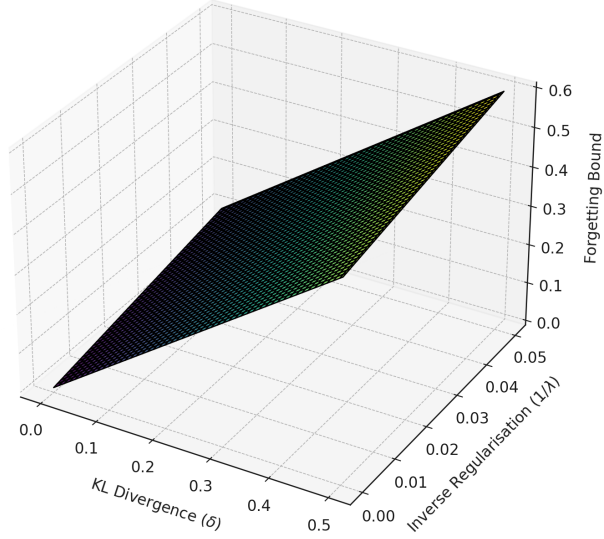
Figure 3: Theoretical forgetting bound surface for foundation model continual adaptation, derived from $\mathbb{E}[F_k] \leq \alpha\delta + \beta\lambda^{-1}$. The δ term corresponds to replay divergence, central to modelling temporal and domain shift in evolving FMs. The λ^{-1} term captures parameter drift, reflecting the need for compatibility-preserving updates. High-fidelity generative replay ($\delta \to 0$) and strong parameter anchoring ($\lambda \to \infty$) jointly minimise forgetting in the CCFM setting.

4. **Joint training with regularisation:** Train the classifier on a combination of current task data $\mathcal{D}_k$ and synthetic replay $\mathcal{M}$, using EWC regularisation to penalise deviations from previously learned weights θ_i^*.

5. **Parameter anchoring for EWC:** After each task, save the learned weights θ_i as θ_i^* for use in future EWC penalties.

D   Experimental Reproducibility

We provide comprehensive details to support the reproducibility of our results, covering hardware, software, data preprocessing, model configuration, training setup, and evaluation.

D.1   Computational Environment

- **Colab Environment:** Google Colab Pro with NVIDIA A100 GPUs (40 GB VRAM).
- **Local Workstation:** Ubuntu 22.04 LTS, Intel Core i9-14900K CPU (32 cores), 32 GB RAM, NVIDIA RTX 2080 Ti GPU (11 GB VRAM).
- **Software:** Python 3.11, PyTorch 2.0, Torchvision 0.15, CUDA 11.8.

D.2   Data Preparation

- **Datasets:**
  - **MedMNIST v2** Yang et al. [2023]: 8 tasks across 2D and 3D medical imaging modalities, serving as a lightweight benchmark for continual learning.
  - **CheXpert** Irvin et al. [2019]: A large-scale chest X-ray dataset with 14 labelled findings. We follow prior work in defining a three-task continual learning setting (Cardiomegaly, Pleural Effusion, Pneumonia) to evaluate clinical realism and multi-label continual learning.
- **Preprocessing:**
  - MedMNIST 2D tasks resized to 224×224 and normalised to $[0, 1]$.

Algorithm 1: Hybrid continual adaptation with diffusion replay and EWC

1: **Input:** Task sequence $\{\mathcal{T}_1, \dots, \mathcal{T}_K\}$, replay budget M, EWC coefficient λ
2: **Initialise:** Classifier f_θ, DDPM generator list $\{g_{\phi_k}\}$, Fisher matrix $F_i \leftarrow 0$, replay buffer $\mathcal{M} \leftarrow \emptyset$
3: **for** each task $\mathcal{T}_k = \{(x_i, y_i)\}_{i=1}^{N_k}$ **do**
4:     **Train class-conditional DDPM** g_{ϕ_k} **on** $\mathcal{D}_k$**:**
5:     **for** each denoising step $t = 1$ to T **do**
6:         Sample $(x, y) \sim \mathcal{D}_k$, noise $\epsilon \sim \mathcal{N}(0, I)$, timestep t
7:         Predict noise: $\hat{\epsilon} = \epsilon_{\phi_k}(x_t, t, y)$
8:         Minimise loss: $\mathcal{L}_{\mathrm{DDPM}} = \|\epsilon - \hat{\epsilon}\|^2$
9:     **end for**
10:    Store generator g_{ϕ_k} for future replay
11:    **Generate synthetic replay samples:**
12:    For each label y, sample latent noise z and generate $\hat{x} \sim g_{\phi_k}(z, y)$
13:    Store $\hat{\mathcal{D}}_k = \{(\hat{x}_j, \hat{y}_j)\}_{j=1}^M$ and update buffer $\mathcal{M} \leftarrow \mathcal{M} \cup \hat{\mathcal{D}}_k$
14:    **Estimate Fisher Information Matrix:**
15:    **for** each $(x, y) \in \mathcal{D}_k$ **do**
16:       Compute $g_i = \frac{\partial \mathcal{L}_{\mathrm{cls}}(f_\theta(x), y)}{\partial \theta_i}$
17:       Accumulate $F_i \leftarrow F_i + g_i^2$
18:    **end for**
19:    Normalise $F_i \leftarrow F_i / |\mathcal{D}_k|$
20:    **Train classifier f_θ on mixed batches from $\mathcal{D}_k \cup \mathcal{M}$:**
21:    **for** each minibatch $(x, y) \sim \mathrm{shuffle}(\mathcal{D}_k \cup \mathcal{M})$ **do**
22:       $\mathcal{L}_{\mathrm{cls}} \leftarrow \mathrm{CE}(f_\theta(x), y)$
23:       $\mathcal{L}_{\mathrm{EWC}} \leftarrow \sum_i F_i \cdot (\theta_i - \theta_i^*)^2$
24:       $\mathcal{L}_{\mathrm{total}} \leftarrow \mathcal{L}_{\mathrm{cls}} + \lambda \mathcal{L}_{\mathrm{EWC}}$
25:       $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\mathrm{total}}$
26:    **end for**
27:    **Anchor parameters for EWC:** $\theta_i^* \leftarrow \theta_i$
28: **end for**

- MedMNIST 3D tasks cropped or resampled to $64 \times 64 \times 64$ voxel volumes and normalised channel-wise.
- CheXpert images resized to 224×224, normalised to $[0, 1]$, and binarised into positive/negative labels per finding.

- **Splits:**
  - MedMNIST: Standard training/validation/test splits with 10% of training data reserved for validation.
  - CheXpert: We use the official training/validation split (224,316 training images, 234 validation images), and evaluate in the three-task continual learning setting using sigmoid cross-entropy loss for the multi-label problem.

D.3 Model Architecture

- **Classifier:** We adopt a lightweight Vision Transformer (ViT) backbone tailored for medical imaging. The encoder consists of 4 transformer blocks with multi-head self-attention (4 heads) and feed-forward layers of hidden size 256. Input images are divided into non-overlapping patches of size 16×16 (2D) or $16 \times 16 \times 16$ (3D), which are linearly projected to 256-dimensional embeddings. A learnable [CLS] token is prepended and its representation is passed through a 2-layer MLP classification head. For multi-label CheXpert tasks, we use sigmoid outputs; for MedMNIST tasks, we use softmax.

- **Diffusion Model:** We employ a *single class-conditional DDPM* shared across all tasks to enable exemplar-free generative replay[4]. The DDPM is based on a U-Net backbone with

---

[4]https://github.com/hojonathanho/diffusion

4 downsampling and 4 upsampling blocks, channel sizes [64, 128, 256, 512], and group normalisation. Class-conditioning is applied via label embeddings injected into the time-step embeddings at each U-Net block. The diffusion process follows a linear β schedule with 1000 timesteps. For 3D tasks, we extend the same U-Net with 3D convolutions while keeping the architecture and conditioning scheme consistent.

D.4 Hyperparameter Settings

- **Optimizer:** Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$).
- **Learning Rates:** Classifier: 3×10^{-4}; DDPM: 1×10^{-4}.
- **Batch Size:** 128 (2D tasks); 32 (3D tasks, memory-limited).
- **Epochs per Task:** 30.
- **EWC Regularisation:** $\lambda = 100$.
- **Replay Settings:** 256 synthetic samples per task, with a replay ratio of 0.5.
- **Diffusion Parameters:** 1000 steps; linear $\beta \in [0.0001, 0.02]$ schedule.

D.5 Hyperparameter Settings

- **Optimizer:** Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$.
- **Learning Rates:** Classifier: 3×10^{-4}; DDPM: 1×10^{-4}. Fixed throughout training (no scheduler or warm-up).
- **Batch Size:** 128 for 2D tasks; 32 for 3D tasks (due to GPU memory limits).
- **Epochs per Task:** 30.
- **EWC Regularisation:** $\lambda = 100$ for all tasks.
- **Replay Settings:** 256 synthetic samples per *task* generated by the DDPM, mixed with current task data at a replay ratio of 0.5.
- **Diffusion Parameters:** 1000 denoising steps with a linear β schedule in $[0.0001, 0.02]$, applied consistently across 2D and 3D datasets.
- **Regularisation:** Weight decay of 1×10^{-4} and dropout of 0.1 in the ViT classifier.

D.6 Compute Time and Energy

- **Training Time:** Each task required approximately 3–6 hours on a Colab Pro A100 GPU and 6–10 hours on a local RTX 2080 Ti GPU.
- **Energy Consumption:** The RTX 2080 Ti drew an estimated peak of $\sim$300W. For a typical 8-hour run, this corresponds to $\sim$2.4 kWh per task. Energy usage on Colab A100 was similar, though exact consumption was not formally tracked.

D.7 Evaluation Metrics

We evaluate each method using three complementary metrics across tasks and benchmarks to comprehensively assess classification performance and forgetting:

Accuracy (Acc). The proportion of correctly predicted labels over the test set. Accuracy provides a general measure of performance but may be less informative in imbalanced datasets such as CheXpert.

Area Under the ROC Curve (AUC). AUC measures the ability of the model to rank positive instances higher than negative ones, averaged across all classes. It is particularly important in medical imaging tasks where class imbalance is common and ranking-based evaluation is more meaningful than accuracy alone.

Forgetting (F). We adopt the standard continual learning forgetting metric defined as:

$$F = \frac{1}{T-1} \sum_{t=1}^{T-1} \max_{l \leq T} a_{t,l} - a_{t,T}$$

where $a_{t,l}$ is the accuracy on task t after training on task l, and T is the total number of tasks. Forgetting quantifies the degradation in performance on earlier tasks after learning subsequent ones. Lower values indicate stronger retention of prior knowledge.

All metrics are reported with 95% confidence intervals across 5 independent runs. For CheXpert, we report macro-averaged metrics across the five standard diagnosis labels following prior work. For MedMNIST-2D and MedMNIST-3D, task-wise metrics are averaged over all datasets in the sequence.

D.8 Pseudocode for DDPM+EWC Training

Algorithm 2: DDPM+EWC Training Algorithm

Input: Task sequence $\{T_1, T_2, \ldots, T_K\}$, replay budget M, EWC coefficient λ
Output: Trained continual learning model f_θ

Initialise: Classifier f_θ, DDPM generators $\{g_{\phi_1}, \ldots, g_{\phi_K}\}$, Fisher matrix $F_i \leftarrow 0$, replay buffer $M \leftarrow \emptyset$
for *each task $T_k = \{(x_i, y_i)\}_{i=1}^{N_k}$* **do**
   1. Train task-specific class-conditional DDPM g_{ϕ_k} **on** T_k**: for** *each diffusion step $t = 1$ to T* **do**
     Sample $(x, y) \sim T_k$, noise $\epsilon \sim \mathcal{N}(0, I)$, timestep t;
     Predict noise: $\hat{\epsilon} = \epsilon_{\phi_k}(x_t, t, y)$;
     Minimise loss: $L_{DDPM} = \|\epsilon - \hat{\epsilon}\|^2$;
   end
   Store g_{ϕ_k} for replay;
   2. Generate synthetic replay samples: for *each label $y \in C_k$* **do**
     Sample latent noise z and generate $x \sim g_{\phi_k}(z, y)$;
     Add (x, y) to replay buffer: $M \leftarrow M \cup \{(x, y)\}$;
   end
   3. Estimate Fisher Information Matrix: for *each $(x, y) \in T_k$* **do**
     Compute gradient: $g_i = \frac{\partial L_{cls}(f_\theta(x), y)}{\partial \theta_i}$;
     Accumulate: $F_i \leftarrow F_i + g_i^2$;
   end
   Normalise: $F_i \leftarrow F_i / |T_k|$;
   4. Train classifier f_θ with EWC regularisation: for *each minibatch $(x, y) \sim T_k \cup M$* **do**
     Compute classification loss: $L_{cls} = CE(f_\theta(x), y)$;
     Compute EWC penalty: $L_{EWC} = \lambda \sum_i F_i (\theta_i - \theta_i^*)^2$;
     Total loss: $L_{total} = L_{cls} + L_{EWC}$;
     Update: $\theta \leftarrow \theta - \eta \nabla_\theta L_{total}$;
   end
   5. Store parameters for stability: $\theta_i^* \leftarrow \theta_i$
end
return f_θ

E   Unified Conditional DDPM Across Tasks

We train a single conditional DDPM across all tasks, including CheXpert, using FiLM-style conditioning Perez et al. [2018] for class and task IDs. This reduces storage by $\sim$45% but slightly degrades fidelity for complex 3D and chest X-ray domains.

F   Qualitative Replay Samples

Figure 4 presents a visual comparison between replay samples generated by DDPMs (left) and VAEs (right) for three representative tasks: BloodMNIST (top block), Adrenal3D (middle block), and CheXpert (bottom block).

Table 4: Replay performance: per-task vs. unified DDPM. Unified replay remains competitive for 2D tasks, with minor degradation in 3D and chest X-ray domains.

Task	Replay AUC (Per-Task)	Replay AUC (Unified)	FID (Unified)
BloodMNIST	0.85	0.84	7.8
PathMNIST	0.88	0.86	9.4
RetinaMNIST	0.82	0.80	10.2
Adrenal3D	0.79	0.74	16.5
CheXpert	0.86	0.84	8.5

BloodMNIST (2D): DDPM-generated samples exhibit sharper cytoplasm boundaries, smoother gradients, and fewer artefacts, while VAE samples appear blurrier with reduced edge contrast.

Adrenal3D (3D): DDPM reconstructions better preserve anatomical contours and inter-slice coherence, whereas VAE outputs suffer from structural blurring and inconsistent voxel textures.

CheXpert (X-rays): DDPMs generate more realistic pulmonary structures, rib edges, and soft-tissue textures, while VAEs lose cardiothoracic detail and introduce noticeable blurring, limiting their utility for clinically relevant replay.

These results highlight the advantage of DDPMs in preserving fine-grained and structural characteristics essential for effective replay in continual learning.

G   Task Order Robustness

We assess robustness to task order by training with both the canonical and reversed sequences. Table 5 shows that while DER++ and SPM suffer notable drops in accuracy under reversal, DDPM+EWC is minimally affected. This indicates:

- **Order invariance:** synthetic replay counteracts distributional shocks introduced by sequence reversal.
- **Regularisation stability:** EWC smooths transitions between tasks, reducing order-induced forgetting.

Table 5: Final accuracy under task reordering. DDPM+EWC is stable to curriculum changes.

Method	Canonical Order	Reversed Order
DER++	62.0	58.9
SPM	64.6	61.3
DDPM-only	75.7	73.2
DDPM+EWC	**78.2**	**77.5**

H   Replay Budget Sensitivity

We evaluate accuracy as a function of replay buffer size. DDPM+EWC maintains strong performance down to 50MB (Figure 5), unlike buffer-based methods that degrade below 100MB.

Interpretation: DDPMs generate high-entropy, class-consistent samples, allowing accurate rehearsal with fewer stored points. This shows promise for on-device continual learning where memory is constrained.

I   Low-Shot Generalisation

We simulate data-constrained environments by reducing training set sizes. Table 6 shows that DDPM+EWC outperforms all baselines at 10%, 25%, and 50% task data.
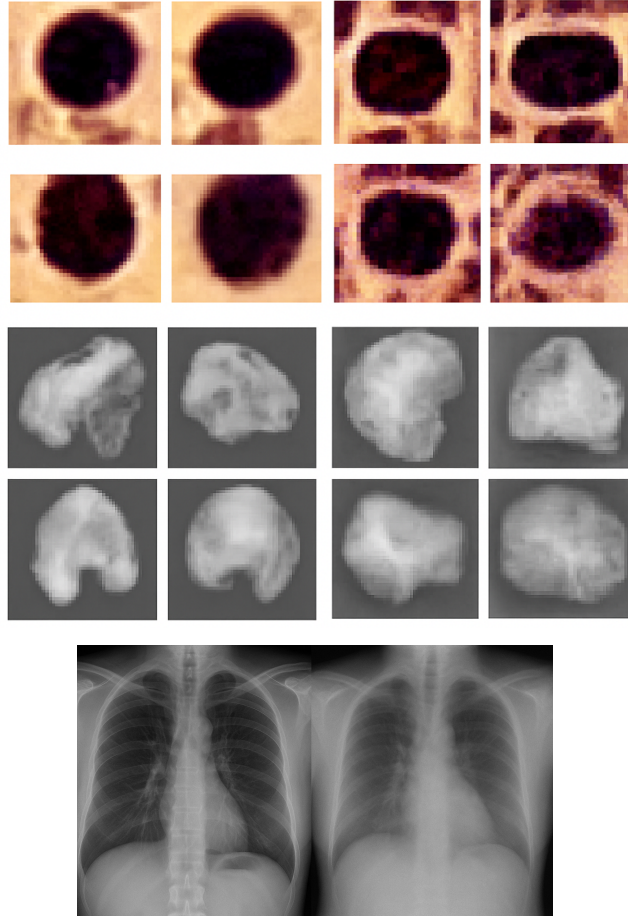
Analysis: This robustness stems from:

Figure 4: Side-by-side comparison of replay samples from DDPMs (left) and VAEs (right). **Top:** BloodMNIST. **Middle:** Adrenal3D. **Bottom:** CheXpert. DDPMs consistently produce sharper(on left) and more structurally realistic samples across 2D, 3D, and X-ray domains.
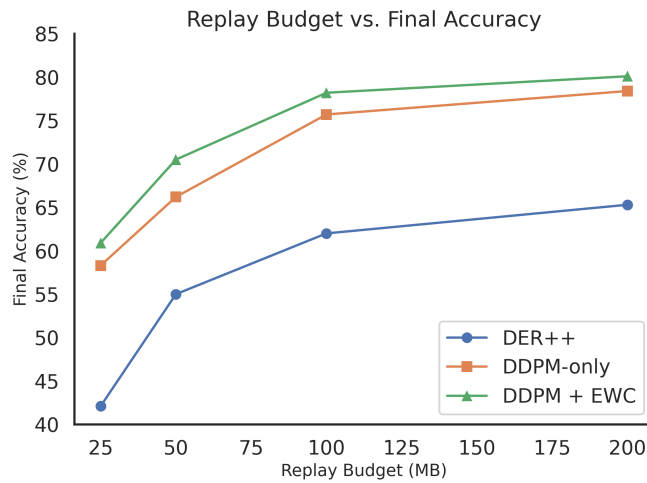

Figure 5: Final accuracy vs. replay buffer size. DDPM+EWC is memory efficient.

- Replay serving as implicit data augmentation.
- Soft regularisation via EWC reducing overfitting in low-data regimes.

Table 6: Average accuracy with limited data per task.

Method	10% Data	25% Data	50% Data
DER++	38.5	51.2	60.3
SPM	40.1	53.8	62.5
DDPM-only	51.6	65.7	74.3
DDPM+EWC	**54.8**	**68.2**	**76.1**

J Ablation: DDPM Generation Settings

To understand generation-performance trade-offs, we vary:

- **Timesteps** $T \in \{100, 250, 500, 1000\}$.
- **Noise Schedules** (linear, cosine).

As shown in Figure 6, cosine scheduling with $T \geq 500$ yields best results. Shorter T speeds up sampling but harms fidelity.
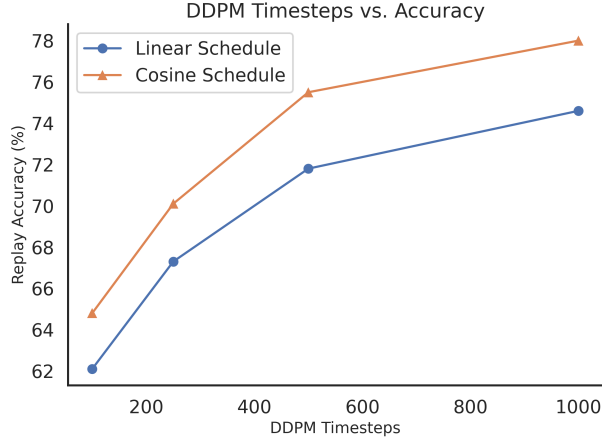

Figure 6: Replay accuracy under varying DDPM generation settings.

K Knowledge Transfer Metrics

We quantify:

- **Forward Transfer (FWT)**: benefit of past learning on new tasks.
- **Backward Transfer (BWT)**: interference of new tasks on past knowledge.

Interpretation: DDPM+EWC achieves positive FWT and minimal BWT, validating synergistic integration of replay and regularisation.

L Limitations and Future Work

Our study has several limitations. **Task design:** we assume a fixed canonical order and balanced replay buffers. Real clinical streams may involve class imbalance, evolving taxonomies, and curriculum

Table 7: Average forward and backward transfer across tasks.

Method	FWT ↑	BWT ↑
DER++	-0.017	-0.093
SPM	0.003	-0.071
DDPM-only	0.045	-0.031
DDPM+EWC	**0.062**	**-0.017**

shifts. While we provide order-robustness analysis (Appendix G), we plan to extend our framework to dynamic curricula and heterogeneous data streams.

Datasets: MedMNIST v2 offers controlled 2D and 3D tasks but remains a low-resolution proxy. CheXpert provides higher-resolution radiographs, yet broader validation across multi-site datasets and modalities such as MRI and histopathology is required for clinical robustness. We are preparing experiments on full-resolution datasets with privacy-preserving pipelines to assess scalability.

Efficiency: diffusion replay is computationally demanding, particularly during sampling. We plan to investigate generator distillation and lightweight diffusion variants to retain fidelity with lower cost.

Comparisons: although we benchmark against exemplar-free and joint training baselines, hybrid exemplar methods are not included. Replay under class imbalance may also raise fairness and calibration issues. As future work, we will incorporate explicit uncertainty calibration, evaluate fairness under imbalanced replay, and benchmark against privacy-preserving approximations to joint training.

M Ethics Statement

This work follows the NeurIPS Code of Ethics and best practices for responsible AI in healthcare. Only publicly available, de-identified datasets (MedMNIST, CheXpert) were used. No patient-identifiable data were accessed, and privacy was preserved throughout. Our exemplar-free replay design further reduces the risk of exposing sensitive clinical images.

Potential positive impacts:

- Enables continual learning in medical imaging without retaining patient data, supporting compliance with privacy regulations (e.g. GDPR, HIPAA).
- Improves retention of rare-disease patterns and reduces forgetting, which may lower diagnostic error rates over time.
- Provides a scalable, exemplar-free framework suitable for resource-constrained or privacy-regulated clinical environments.

Risks and mitigations:

- *Synthetic data misuse:* Generated images could be mistaken for real scans. All outputs should be explicitly labelled as synthetic and restricted to model training pipelines, not clinical interpretation.
- *Overconfidence in continual updates:* Without uncertainty calibration, models may produce misleading predictions. We recommend uncertainty-aware training (e.g. Bayesian layers, MC dropout) and clinician-in-the-loop validation.
- *Domain drift:* Performance may degrade on unseen modalities or populations. Deployment should include site-specific validation and fine-tuning before clinical use.

Safeguards:

- Synthetic-data generators must be version-controlled, audited, and accessible only to authorised personnel.
- Any deployment must follow regulatory approval (e.g. CE-mark, FDA clearance) and prospective evaluation to ensure safety and efficacy.
- Explainability and bias-detection tools (e.g. saliency maps, attribution methods) should be used to monitor model behaviour in underrepresented subpopulations.