Conditional Diffusion Replay for Continual Learning in Medical Settings

Yewon Byun [1] **Saurabh Garg** [1] **Sanket Vaibhav Mehta** [1] **Praveer Singh** [2 3] **Jayashree Kalpathy-Cramer** [2 3]
Bryan Wilder [1] **Zachary Lipton** [1]

Abstract

Episodic replay methods, which store and replay past data, have proven effective for handling distribution shifts in continual learning. However, due to regulatory and privacy concerns for data sharing, their applicability can be limited, especially in healthcare. In this work, we advance the state of art, focusing our inquiry on two novel benchmarks for domain incremental continual learning: diabetic retinopathy severity classification and dermoscopy skin lesion detection. First, we demonstrate the poor forward and backward transferability of simple baselines. Then, to overcome these challenges, we propose a novel method called *conditional diffusion replay*. By leveraging a text-to-image diffusion model for synthetic data generation, our approach effectively preserves performance on previously encountered domains while adapting to new ones. We observe that compared to standard sequential fine-tuning, our conditional diffusion replay method improves average AUC by up to 7.3 points and 3.3 points for the skin lesions and diabetic retinopathy benchmarks, respectively.

1. Introduction

Machine learning models deployed in real-world settings frequently encounter performance degradation due to distribution shifts, which are characterized by statistical differences between the data encountered during deployment and the data used in training (Quinonero-Candela et al., 2008; Torralba & Efros, 2011; Koh et al., 2021; Garg et al., 2022; 2023). This issue is particularly prevalent in healthcare, where disparities across various factors can lead to significant drops in out-of-distribution performance. These factors include patient demographics, institutions, types of scanners, image acquisition techniques, and both inter and intra-rater variability in annotated labels (Wantlin et al., 2023). In such cases, given access to labeled data from drifting distributions, the model should be able to adapt to new distributions (i.e., *forward transfer*), while also preserving its performance on previously encountered distributions (i.e., *backward transfer*). Continually training the model on drifting distributions is one basic approach. However, this can result in failures of backward transfer (sometimes referred to as *catastrophic forgetting* (McCloskey & Cohen, 1989)) where the acquisition of new tasks leads to the erasure of previously acquired knowledge.

One simple solution might be to gather all data in a single central location and train on the amalgamation of all distributions. However, this approach is often not viable in real-world scenarios like healthcare, where regulatory concerns, privacy constraints, security requirements, intellectual property considerations, and general unease about sharing data can create formidable obstacles to sharing among facilities. Even continual learning (CL) approaches including episodic replay-based methods, which have recently shown promising results (Chaudhry et al., 2019), can run afoul of the aforementioned concerns. These methods store past data in a replay buffer and periodically sample from the buffer to mitigate catastrophic forgetting. However, for application in medical settings, these methods run into the same limitations owing to the need to share data across settings. Several alternative approaches like recent prompt-based methods (Wang et al., 2022b; Smith et al., 2023b) sidestep the need for storing data and nevertheless have shown strong performance on standard CL benchmarks. However, as our experiments bear out, it can be difficult to apply these methods successfully to real-world problems (see Section 5.3).

In this work, we tackle distribution shift problems in two *novel* healthcare benchmarks for CL: diabetic retinopathy severity classification and dermoscopy lesion detection (Wantlin et al., 2023). We demonstrate that naive sequential fine-tuning for incremental domains presents issues with poor backward transferability. To mitigate these problems and improve backward transfer, our method leverages the power of generative models. Our key insight is that in a world in which model sharing is acceptable but data shar-

[1]School of Computer Science, Carnegie Mellon University [2]Department of Radiology, Massachusetts General Hospital [3]Department of Ophthalmology, University of Colorado. Correspondence to: Yewon Byun <yewonb@cs.cmu.edu>.

Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

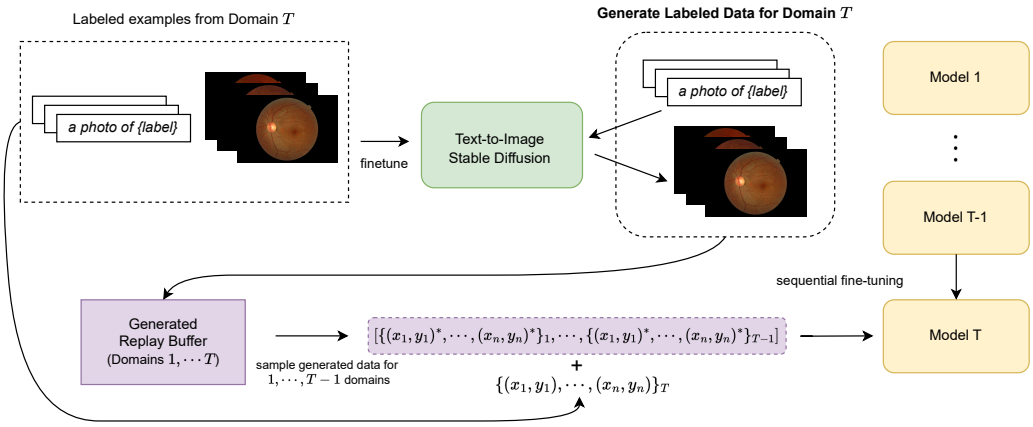

Figure 1. **Conditional Diffusion Replay**. First, we fine-tune an off-the-shelf text-to-image stable diffusion model, with image-caption pairs from domain T. Captions are of form *a photo of* $\mathit{flabelg}$. As an example, in the context of Diabetic Retinopathy, $\mathit{flabelg}$ can correspond to *Proliferative Diabetic Retinopathy*. Next, we sample from our fine-tuned diffusion model, by inputting text captions. This allows us to generate labeled data for domain T, which we then store in our replay buffer. We then sample generated data for previous domains $1, ..., T$ 1 from our generated replay buffer. We combine this with our real data from our current domain T. Lastly, we fine-tune our model on the union of real data of our current domain and generated data of previous domains.

ing is not, we could share not only discriminative models across domains but also generative models. In our approach, we generate synthetic data to simulate previous domains while training on each incremental domain. Our instantiation of this approach, *conditional diffusion replay* (see Figure 1), leverages recent advances in stable diffusion models to create realistic images from textual descriptions (Rombach et al., 2022).

In medical imaging contexts, well-employed techniques to capture novel concepts, such as textual inversion methods (Gal et al., 2022) are inapplicable, since they presume that the diffusion model contains the necessary information to control the generation process for novel words. To imbue the generator with new concepts, we fine-tune a pre-trained stable diffusion model on novel concepts using a limited number of samples, generating high-quality images that support downstream discriminative tasks. Our conditional diffusion replay-based method seeks to preserve and reinforce previously learned knowledge using *generated labeled data* while adapting to new data distributions. The effectiveness of this approach, especially in small-scale, real-world datasets, is not readily apparent. However, we empirically demonstrate its superiority over state-of-the-art methods on both benchmarks. Our approach paves the way for CL in sensitive domains, promoting the development of resilient, adaptive models while maintaining the privacy and ethical considerations. In summary, the primary contributions of this paper include:

- To overcome the limitation of CL benchmarks primarily focused on synthetic and simplified settings, we present two new benchmarks derived from real-world healthcare data. These benchmarks serve as valuable tools for assessing the robustness of CL algorithms in real-world scenarios, particularly in the domain-incremental setting.
- We propose a novel method, *Conditional Diffusion Replay*, for effective adaptation to evolving data distributions, by fine-tuning and sampling from a text-to-image diffusion model.
- We observe that compared to sequential fine-tuning, our method improves average AUC by up to 7.3 and 3.3 points for the skin lesions and diabetic retinopathy benchmarks, respectively, and also outperforms state-of-the-art methods.

2. Preliminaries

2.1. Problem Setup: Domain-Incremental Learning

In this paper, we focus on the *domain-incremental scenario* in CL, where the primary goal is to learn a model that adapts to a new domain while alleviating catastrophic forgetting on previously seen domains. Formally, we consider a sequence of T domains, D_1 ! ! D_T, where $D_t = fx_i^t, y_i^t g_{i=0}^{N_t}$ represents a dataset corresponding to domain t, sampled from underlying distribution $P_t(X, Y)$, x_i^t 2 X is the i-th image with y_i^t 2 Y as its label and N_t is the total number of samples for domain t. Furthermore, in the domain incremental scenario, $P_t(X)$ 6= $P_{t+1}(X)$ and $P_t(Y)$ 6= $P_{t+1}(Y)$, 8t, while label space Y remain fixed across all domains. The goal is to learn a predictor $f_\theta : X$! Y, parameterized by θ 2 R^P, to minimize the average expected risk of all N domains. To demonstrate the model's learning behavior over

the sequence of domains and analyze catastrophic forgetting of the previously seen domains, we evaluate the model after training on a specific domain t using the test dataset of that domain, D_t^{test} $P_t(X, Y)$, and from past domains, D_i^{test} $P_i(X, Y), \forall i$, where $1 \quad i \quad t \quad 1$. During sequential training, the domain identity is known. However, during inference, the model must implicitly determine the domain identifier before making predictions. This makes the domain-incremental scenario more challenging than a task-incremental scenario, where the identifier is known even during inference (Van de Ven & Tolias, 2019).

2.2. Evaluation Metrics

Let $\alpha_{s,t}$ denote the AUC on domain s after training on domain t. To evaluate the model's robustness against catastrophic forgetting after training on domain t, we report the following metrics: **average AUC** (A_t), **forgetting** (F_t), and **learning AUC** (LA_t) (Lopez-Paz & Ranzato, 2017; Riemer et al., 2019; Mehta et al., 2021). We note that the original metrics in the paper are in terms of accuracy (i.e. average *accuracy*, learning *accuracy*). However, we choose to compute these metrics in terms of AUC given the presence of label imbalance in our tasks. F_t measures *backward transfer*: how does learning on domain t influences the performance of previous domains $s, 1 \quad s < t$. LA_t indirectly measures *forward transfer*: the learning capability when the model encounters new domain t. We can formally define the above metrics as follows:

$$A_t = \frac{1}{t}\sum_{s=1}^{t} \alpha_{s,t}; LA_t = \frac{1}{t}\sum_{s=1}^{t} \alpha_{s,s} \tag{1}$$

$$F_t = \frac{1}{t \quad 1}\sum_{s=1}^{t \quad 1} \max_{s' 2 f 1,\ldots,t \quad 1g}(\alpha_{s,s'} \quad \alpha_{s,t}) \tag{2}$$

2.3. Baseline Comparisons

First, we compare with the naive finetune (FT) baseline, which sequentially finetunes the model on each domain without additional constraints in the learning process. Elastic weight consolidation (EWC) (Kirkpatrick et al., 2017) is a regularization-based approach that constrains parameters to lie in regions of low error for previous domains, by applying a penalty term determined by the Fisher information matrix.

Another family of CL methods assumes the presence of an episodic replay (ER) buffer, storing a subset of data from previous domains (Sodhani et al., 2022). ER methods have demonstrated state-of-the-art performance on existing benchmarks, with even small buffer sizes (Chaudhry et al., 2019; Wang et al., 2020). Intuitively, the oracle performance in the CL setup would be training on the union of $fD_1, ..., D_t g$, which is why having access to samples from previous domains directly alleviates forgetting to some extent. While replay-based methods have demonstrated competitive performance, they rely on access to previous data. This is not an ideal solution for many real-world scenarios, such as healthcare, where data sharing is prohibited due to privacy and regulatory constraints. To address the issue of avoiding the storage of real samples from previous domains, Deep Generative Replay (DGR) (Shin et al., 2017), utilizes a generative model trained within the framework of Generative Adversarial Networks (GANs) to produce "unlabeled data" that resembles the distributions of previous domains and pseudo-labels the generated samples. However, we show in Section 5.2, that even after replacing WC-GAN with Stable Diffusion (DGR++) for a fair comparison, the approach suffers from generating samples for previous domains that follow the true label distribution.

In response to the popularity of pre-trained models, prompt-based continual learning approaches like DualPrompt (Wang et al., 2022b) and CODA-Prompt (Smith et al., 2023b) have been introduced. These methods involve learning a small number of parameters per domain using continuous token embeddings or prompts while keeping the rest of the pre-trained model frozen. The appropriate prompt is selected based on the input data. However, these methods rely on domain-specific pre-trained models, which may not be accessible in sensitive environments like healthcare. Further, these methods represent prompts as learnable continuous tokens while we consider prompts as discrete text tokens.

3. Conditional Diffusion Replay

To overcome the limitations of past generative replay methods (Zhao et al., 2022; Van de Ven et al., 2020; Lesort et al., 2019; Shin et al., 2017), we introduce our method: *Conditional Diffusion Replay*, which seeks to achieve oracle performance by learning to generate *labeled data* that resemble the underlying distribution of previously seen domains. Thus, without having to store true data from previous domains, we can retain some representation of past distributions conditioned on the quality of our generations.

3.1. Algorithm

At each domain t, we do the following: First, we recall generated samples of past domains $M_1, ..., M_{t \quad 1}$ from our replay buffer and finetune our classifier f_{θ_t} on the union of generated and real samples $M_1 [... [M_{t \quad 1} [f(x_i^t, y_i^t) g_{i=0}^{N_t}$. Second, we finetune a text-to-image generator (e.g., Stable Diffusion; Rombach et al. (2022)), G_t, using the dataset from domain t (D_t), then generate N_t new samples from G_t and store them to M_t. The proposed method is outlined in Algorithm 1 (see Appendix A).

We finetune our generator G_t with conditional diffusion, where the model is conditioned on textual prompts. We do this specifically to 1) generate labeled data and 2) control

Method	Skin Lesions						Diabetic Retinopathy					
	Avg AUC ″		Forgetting #		Learning AUC ″		Avg AUC ″		Forgetting #		Learning AUC ″	
FT	82.67	1.83	14.46	2.39	92.31	0.31	78.49	0.96	12.6	2.1	84.79	0.13
EWC	84.56	0.77	11.97	0.91	92.54	0.20	76.80	0.24	15.63	0.41	84.61	0.15
DGR ++	78.27	2.25	17.43	2.79	89.89	1.36	66.89	1.12	3.14	0.10	68.46	1.17
DualPrompt	84.59	0.38	8.67	0.64	90.37	0.17	80.89	0.30	7.90	0.36	84.85	0.21
CODA-Prompt	86.85	0.77	9.82	1.05	93.39	0.08	81.73	1.13	**3.07**	**1.59**	83.23	1.93
Conditional Diffusion Replay (CDR)	**89.92**	**0.55**	**5.02**	**0.79**	93.27	0.06	**81.76**	**0.27**	6.19	0.42	84.85	0.27
Oracle Performance	92.65	0.17	0.57	0.21	92.77	0.18	84.11	0.29	0.7	0.57	84.38	0.28

Table 1. Comparing performance in terms of average AUC, forgetting, and learning AUC across methods after training on the last domains. ″ indicates higher is better, # indicates lower is better. For the skin lesions benchmark, we average over all possible six domain sequences. For the diabetic retinopathy benchmark, we evaluate the performance of the Messidor *!* APTOS sequence, averaging over 3 random seeds. We note that we only evaluate the unilateral direction for this setting, due to no performance drop in the APTOS *!* Messidor sequence. We observe that compared to vanilla FT, Conditional Diffusion Replay improves average AUC by up to 7.3 points and 3.3 points for the skin lesions and diabetic retinopathy benchmarks, respectively, and also outperforms state-of-the-art methods.

for the label distribution $P_t(Y)$. Since the generator is conditioned on textual prompts, at inference time, we can explicitly control for $P_t(Y)$ to correctly represent the true label distribution. DGR uses a classifier f_{θ_t} to create pseudo-labels for the generated unlabeled samples. As the number of domains increases, employing this approach is likely to lead to a compounded error. Furthermore, it is unclear how these existing methods, which have only been evaluated on balanced synthetic datasets (i.e., Rotated MNIST), behave in the setting of high label imbalance, as reflected in many healthcare datasets (Wantlin et al., 2023). In this setting, it is unclear if the generator will be able to mimic the data distribution or if it will collapse onto the majority class, leading to a skewed representation of past distributions.

We note that the performance of our method is heavily reliant on the quality of the generated images. Thus, improving the finetuning procedure of stable diffusion will lead to better downstream performance and constitutes future work.

4. Datasets and Task Sequences

We propose two new domain-incremental CL benchmarks – *Diabetic Retinopathy* and *Skin Lesions*, built using BenchMD (Wantlin et al., 2023). The diabetic retinopathy (DR) benchmark consists of two domains of 2D eye fundus images (Messidor-2, APTOS), for a multi-class classification task, over 5 unified labels of the severity of DR. The Skin Lesions benchmark consists of three domains of dermoscopic image datasets (BCN2000, PAD-UEFS-20, and HAM10000), for a multi-class classification task, over 5 unified labels of skin lesions. Distribution shifts between domains in both datasets exist in terms of demographics, collection period, camera types, and image quality. For details, please refer to Appendix C.

5. Experimental Results

Table 3 reports the results of our method (CDR) compared to various baselines, including standard sequential fine-tuning (FT), elastic weighted consolidation (EWC), DualPrompt, CODA-Prompt, and our adapted version of deep generative replay (DGR) with Stable Diffusion, termed DGR++. For details, please refer to Appendix D.

5.1. How does CDR perform on realistic benchmarks?

We observe that CDR significantly outperforms FT for both real-world healthcare benchmarks, showcasing the superiority of our method for alleviating forgetting. Concretely, CDR improves the average AUC over FT by 7.25 and 3.27 on skin lesions and diabetic retinopathy benchmarks, respectively. Additionally, we present results for the scenario where access to all previous data is assumed, referred to as Oracle Performance. This enables us to establish an upper-bound reference point for this setup. Upon examining Table 3, we note an average AUC performance decrease of approximately 3 points compared to the Oracle upper-bound. These results emphasize the need for future investigations on these benchmarks to address the performance drop and further enhance the methodology.

5.2. How does CDR perform in comparison to DGR++?

Despite both belonging to the category of generative replay-based approaches for continual learning, we observe a significant performance advantage of CDR over DGR++ based on the results presented in Table 3. One of the main advantages of our text-conditioned generation approach is that it allows us to generate samples while following the ground-truth label distribution. In contrast, DGR++ generates unlabeled samples and assigns pseudo labels using a classifier,

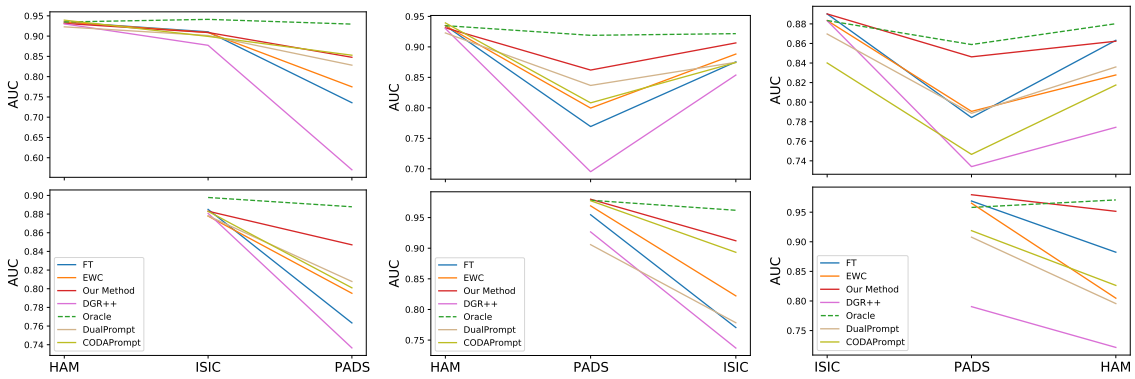

Figure 2. We visualize how the test AUC evolves throughout three distinct domain sequences for domain 1 (top) and domain 2 (bottom) for the skin lesions benchmark. As an example, 1) the upper left plot shows the model performance throughout the fine-tuning sequence (i.e. HAM-ISIC-PADS), evaluated on Domain 1 (i.e. HAM) test data 2) the bottom left plot shows the model performance, evaluated on Domain 2 (i.e. ISIC) test data. We also provide the results of the *Oracle Performance*, where all past task data is assumed to be available, representing the highest achievable performance in our scenario. CDR shows a significant enhancement in performance compared to other baseline methods. For additional domain sequences, please refer to Figure 3 (Appendix B).

which lacks explicit control over the label distribution. This limitation becomes particularly problematic when working with real-world datasets that exhibit high label imbalance.

Additionally, DGR++ relies on an imperfect labeling function, continuously updating the model with noisy samples as the task sequence length increases. This process compounds errors and is further exacerbated by artifacts present in the generated images. Lastly, DGR shares the generative model across domains, while our method shares the generated samples, to avoid catastrophic forgetting within the diffusion model.

5.3. How does CDR perform in comparison to recent, prompt-based approaches for continual learning?

Prompt-based continual learning approaches, such as DualPrompt and CODA-Prompt, utilize domain-specific prompts and select the appropriate prompt based on the input data during inference. We find that CDR performs 1) comparably to the state-of-the-art CODA-Prompt method for the diabetic retinopathy benchmark, and 2) significantly outperforms CODA-Prompt for the skin lesions benchmark. We acknowledge that prompt-based methods implicitly assume similarity between the pre-trained models and the task-of-interest, which may not hold in real-world scenarios. For instance, pre-trained vision models on ImageNet may struggle to efficiently learn medical tasks in the healthcare domain (see learning AUC for DualPrompt on skin lesions in Table 3).

5.4. How robust is CDR to domain ordering?

To assess the robustness of CDR to the ordering of domains, we conduct experiments by varying the sequence in which the domains are encountered. We visualize the performance drop across all six different domain orderings of the Skin Lesions benchmark, as shown in Figure 2 (refer to Figure 3 in Appendix B for the remaining plots). We observe that in most cases, CDR outperforms considered baselines across the six domain orderings, supporting its practical applicability in real-world scenarios. We note that carefully inspecting how performance evolves throughout the task sequence, uncovers much more drastic differences in performance than what is captured by average AUC and forgetting metrics.

6. Discussion

In this paper, we investigate performance degradation in the presence of real-world distribution shifts encountered in a sequential domain-incremental setting. Our study focuses on what is, to the best of our knowledge, the first use case of text-to-image Stable Diffusion to emulate an episodic replay buffer, addressing these shifts effectively. We demonstrate the effectiveness of our approach by outperforming state-of-the-art baselines on real-world healthcare benchmarks. Our method has the potential to greatly benefit real-world scenarios with limited-labeled data, as they often encounter continuous distribution changes over time.

7. Acknowledgements

We thank Aarushi Gupta and Sharut Gupta for initial discussion on the project. We gratefully acknowledge the NSF (FAI 2040929 and IIS2211955), UPMC, Highmark Health, Abridge, Ford Research, Mozilla, PwC Center, Amazon AI, JP Morgan Chase, Block Center, the Center for Machine Learning and Health, and the CMU Software Engineering Institute via Department of Defense contract FA8702-15-D-0002, for their generous support of ACMI Lab's research.

References

Baweja, C., Glocker, B., and Kamnitsas, K. Towards continual learning in medical imaging. *CoRR*, abs/1811.02496, 2018. URL http://arxiv.org/abs/1811.02496.

Chaudhry, A., Ranzato, M., Rohrbach, M., and Elhoseiny, M. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*.

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

de Masson D'Autume, C., Ruder, S., Kong, L., and Yogatama, D. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis, 2021.

Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D'Amour, A., Moldovan, D., et al. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16458–16468, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Farquhar, S. and Gal, Y. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-Or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

Garg, S., Balakrishnan, S., Lipton, Z., Neyshabur, B., and Sedghi, H. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations (ICLR)*, 2022.

Garg, S., Erickson, N., Sharpnack, J., Smola, A., Balakrishnan, S., and Lipton, Z. Rlsbench: A large-scale empirical study of domain adaptation under relaxed label shift. In *International Conference on Machine Learning (ICML)*, 2023.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks, 2014.

Guan, H. and Liu, M. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

Guo, Y., Liu, M., Yang, T., and Rosing, T. Improved schemes for episodic memory-based lifelong learning. *Advances in Neural Information Processing Systems*, 33: 1023–1035, 2020.

Gupta, S., Singh, P., Chang, K., Qu, L., Aggarwal, M., Arun, N., Vaswani, A., Raghavan, S., Agarwal, V., Gidwani, M., et al. Addressing catastrophic forgetting for medical domain expansion. *arXiv preprint arXiv:2103.13511*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020.

Karani, N., Chaitanya, K., Baumgartner, C., and Konukoglu, E. A lifelong learning approach to brain mr segmentation across scanners and protocols. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pp. 476–484. Springer, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2022.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, mar 2017. doi: 10.1073/pnas.1611835114. URL https://doi.org/10.1073%2Fpnas.1611835114.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Lesort, T., Gepperth, A., Stoian, A., and Filliat, D. Marginal replay vs conditional replay for continual learning. In *International Conference on Artificial Neural Networks*, pp. 466–480. Springer, 2019.

Lopez-Paz, D. and Ranzato, M. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Mehta, S. V., Patil, D., Chandar, S., and Strubell, E. An empirical investigation of the role of pre-training in lifelong learning. *arXiv preprint arXiv:2112.09153*, 2021.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2010.

Perkonigg, M., Hofmanninger, J., Herold, C. J., Brink, J. A., Pianykh, O., Prosch, H., and Langs, G. Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. *Nature communications*, 12(1): 5678, 2021.

Pooch, E. H., Ballester, P., and Barros, R. C. Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In *Thoracic Image Analysis: Second International Workshop, TIA 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2*, pp. 74–83. Springer, 2020.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. Mit Press, 2008.

Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., , and Tesauro, G. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1gTShAct7.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. URL https://arxiv.org/abs/2112.10752.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Shin, H., Lee, J. K., Kim, J., and Kim, J. Continual learning with deep generative replay. *CoRR*, abs/1705.08690, 2017. URL http://arxiv.org/abs/1705.08690.

Smith, J. S., Hsu, Y.-C., Zhang, L., Hua, T., Kira, Z., Shen, Y., and Jin, H. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023a.

Smith, J. S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., and Kira, Z. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11909–11919, 2023b.

Sodhani, S., Faramarzi, M., Mehta, S. V., Malviya, P., Abdelsalam, M., Janarthanan, J., and Chandar, S. An introduction to lifelong supervised learning. *arXiv preprint arXiv:2207.04354*, 2022.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Van de Ven, G. M. and Tolias, A. S. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

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

Wang, Y., Huang, Z., and Hong, X. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=ZVe_WeMold.

Wang, Z., Mehta, S. V., Poczós, B., and Carbonell, J. G. Efficient meta lifelong-learning with limited memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 535–548, 2020.

Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.-Y., Ren, X., Su, G., Perot, V., Dy, J., et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pp. 631–648. Springer, 2022b.

Wang, Z., Zhang, Z., Lee, C.-Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., and Pfister, T. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022c.

Wantlin, K., Wu, C., Huang, S.-C., Banerjee, O., Dadabhoy, F., Mehta, V. V., Han, R. W., Cao, F., Narayan, R. R.,

Colak, E., et al. Benchmd: A benchmark for modality-agnostic learning on medical images and sensors. *arXiv preprint arXiv:2304.08486*, 2023.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. PMLR, 2017.

Zhao, Y., Zheng, Y., Tian, Z., Gao, C., Yu, B., Yu, H., Li, Y., Sun, J., and Zhang, N. L. Prompt conditioned vae: Enhancing generative replay for lifelong learning in task-oriented dialogue. *arXiv preprint arXiv:2210.07783*, 2022.

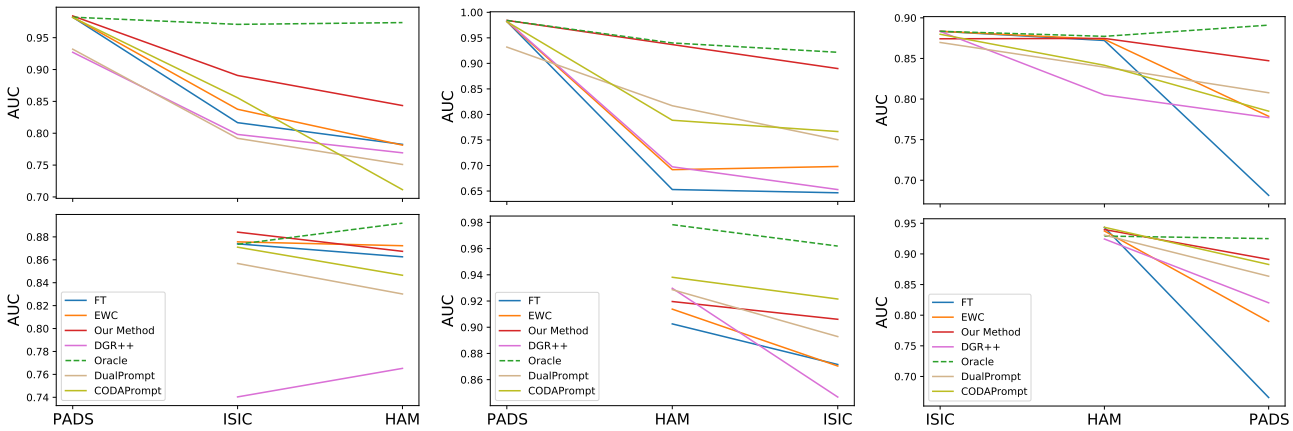

Figure 3. We report how the hold-out test performance evolves throughout the task sequence. This figure contains results for the three remaining task sequences from the Skin Lesions benchmark.

A. Algorithm

Algorithm 1 Conditional Diffusion Replay (CDR)

Procedure CONTINUAL_TRAIN
Input: θ_0, M $fg, D_1 \dots D_T$
Output: θ_T, M
for t $1 \dots T$ **do**
  θ_t $\theta_{t\ 1}$
  for k $1 \dots jD_tj$ **do**
    b_t^k $D_t[k]$ fSample a mini-batch for current domain tg
    r_t^k M fSample a mini-batch for previous domains from the generated replay buffer$g$
    θ_t^{k+1} UPDATE_MODEL$(\theta_t^k, b_t^k, r_t^k)$
  end for
  M UPDATE_MEMORY(M, D_t)
end for
Procedure UPDATE_MEMORY
Input: M, D_t
Output: M
G_t Train Text-to-Image Diffusion Model (D_t)
M_t Sample jD_tj examples from G_t
M M [M_t

B. Additional Results.

We include the remaining 3 task sequences of the Skin Lesions benchmark in Figure 3, to demonstrate the robustness of our method against different orderings.

C. Dataset Details

Our benchmarks are built on top of BenchMD (Wantlin et al., 2023), a recently proposed medical benchmark to study out-of-distribution performance in medical image and sensor data. For our setting, we modify the datasets to have balanced samples per domain. We ensure to preserve the original label distribution of the dataset. Thus, we work in the more challenging and realistic setting of label imbalance (within each domain) and label shifts across domains, i.e., $P_0(Y) \neq \dots \neq P_{T\ 1}(Y) \neq P_T(Y)$.

C.1. Diabetic Retinopathy Benchmark

Messidor-2 consists of high-quality retinal images, collected from French institutions between 2004 to 2010 and APTOS contains more variability in data quality, collected from the Aravind Eye Care System in India, with an unknown collection period. Non-mydriatic cameras were used to collect Messidor-2, while both mydriatic and non-mydriatic cameras were used to collect APTOS. We note that there is no performance drop for the task sequence APTOS ! Messidor-2. Thus, for this setting, we study the Messidor ! APTOS task sequence.

C.2. Skin Lesions Benchmark

Conducting classification on dermoscopic images presents complexities arising from intraclass variations encompassing lesion texture, scale, and color. This benchmark offers a total of six, distinct task sequences. Distribution shifts are present across all three domains (BCN2000, PAD-UEFS-20, and HAM10000), in both demographics and data collection techniques. BCN2000 dataset was collected from Spanish hospitals between 2010 and 2016, PAD-UEFS-20 dataset was obtained from Brazilian hospitals in 2020, and HAM10000 dataset was gathered over the past 20 years from hospitals in Austria and Australia. Dermatoscopes were used for collecting images in BCN2000 and HAM10000, while smartphone cameras were utilized for PAD-UFES-20.

D. Implementation Details.

In our method, we use Stable Diffusion (Rombach et al., 2022) as our conditional diffusion model, with weights from the stabilityai/stable-diffusion-2-1-base checkpoint. For the downstream classification task, we use a ViT-B/16 backbone (Dosovitskiy et al., 2020), pretrained on ImageNet-1K (Russakovsky et al., 2015). We implement baselines with the same architecture and pretrained checkpoint, for consistency. For hyperparameter search, we use the source hold-out performance to select the best combination of parameters. For each dataset, we perform a sweep over different combinations of learning rate α 2 $[1e \quad 3, 1e \quad 2, 5e \quad 2, 0.01]$, and batch size 2 $[32, 64, 128]$. We keep the chosen hyperparameters fixed throughout the entire task sequence. To ensure that we are evaluating our baselines comprehensively, we also run hyperparameter search for different regularization values λ 2 $[0.5, 1, 10, 100]$ for the Elastic Weight Consolidation (EWC) method.

E. Related Work.

Distribution Shifts in Healthcare. Clinical deployment of deep learning models at large scale warrants robustness and effective generalizability of these models on unseen domains. Medical datasets are prone to changes both within as well as across healthcare institutions. Internally, hospitals tend to update their devices, imaging protocols, policies and workflows from time to time. While traversing across institutions, we see changes in the patient profiles, clinical decisions as well as scanner types. This ultimately results in distributional shifts across time and space, thus leading to failure of medical models tested on external testsets. Brittleness of deep learning models in healthcare settings has been a key concern (Zech et al., 2018; Pooch et al., 2020; Guan & Liu, 2021), where models trained on one dataset tend to fail miserably when encountered with dataset shifts. Our work is closest to (Perkonigg et al., 2021; Karani et al., 2018; Gupta et al., 2021; Baweja et al., 2018) which have developed continuous learning techniques for medical imaging, to account for such distribution shifts.

Finetuning for OOD evaluation. In a distinct line of research, numerous studies have probed various strategies for fine-tuning pre-trained models to maintain robust performance when assessed Out-of-Distribution (OOD), that is, on a test distribution that deviates from the fine-tuning data distribution (Pan & Yang, 2010; Djolonga et al., 2021). The main objective of these papers is to enhance the transferability of the fine-tuned models without needing any labeled or unlabeled data from the test distribution. Crucially, since we do get access to labeled data from shifting distributions in a sequential manner, our setup differs from conventional transfer learning problems and resembles more with continual learning setups, as discussed next.

Continual Learning. In this work, we focus on the domain-incremental scenario of continual learning (Van de Ven & Tolias, 2019). In this scenario, the learning system must adapt to new domains while retaining knowledge of previously learned domains. Several methods have been proposed to address forgetting in this scenario, and they can be categorized into different categories (Sodhani et al., 2022). The first category of methods revolves around parameter-based regularization techniques. Two notable methods within this category are Elastic Weight Consolidation (EWC; Kirkpatrick et al. (2017)) and Synaptic Intelligence (SI; Zenke et al. (2017)). Both EWC and SI assess the importance of parameters related to previous

domains and utilize a penalty term to safeguard the knowledge stored in those parameters while updating them for new domains.

Another set of methods focuses on data-based regularization techniques. These approaches retain a subset of data from previous domains as an episodic memory, which is sparsely replayed during the learning of new domains. Several replay-based methods have been proposed, each differing in whether the episodic memory is utilized during training, such as GEM (Lopez-Paz & Ranzato, 2017), A-GEM (Chaudhry et al.), ER (Chaudhry et al., 2019), MEGA (Guo et al., 2020), or during inference, like MbPA (de Masson D'Autume et al., 2019; Wang et al., 2020). These methods assume that the true data can be retained for replay. However, they become less effective in applications where privacy is essential, such as medical settings. To address this limitation, deep generative replay-based methods have been introduced (DGR; Shin et al. (2017)). The main idea behind these methods is to learn a generative model of the dataset and use it to generate samples for experience replay. Additionally, there have been recent works investigating conditional generative replay methods, using GAN-based and VAE architectures (Van de Ven et al., 2020; Zhao et al., 2022; Lesort et al., 2019).

In response to the increasing popularity of pre-trained models, another approach has emerged in the field of continual learning. Mehta et al. (2021) demonstrate that pre-trained initializations implicitly mitigate the issue of forgetting when sequentially fine-tuning models. Another line of approaches, known as prompt-based continual learning, exemplified by L2P (Wang et al., 2022c), DualPrompt (Wang et al., 2022b), S-Prompt (Wang et al., 2022a), and CODA-Prompt (Smith et al., 2023b), involves learning a small number of parameters per domain in the form of continuous token embeddings or prompts while keeping the remaining pre-trained model fixed. The appropriate prompt is then selected based on the input data. Although these methods allow for continual learning without rehearsal, they depend on access to generic pre-trained models, which may not be available in sensitive environments such as healthcare. In all of these methods, prompts are represented as learnable continuous tokens, whereas we consider text prompts represented by discrete tokens.

Various benchmarks, such as Permuted MNIST, Rotated MNIST, and Split CIFAR-100, have been utilized in previous studies to develop algorithms that mitigate forgetting. However, Farquhar & Gal (2018) shows that the phenomenon of catastrophic forgetting is not observed in synthetic Permuted MNIST, casting doubt on the usefulness of such benchmarks. The continual learning community recognizes the importance of challenging and realistic benchmarks in the literature (Farquhar & Gal, 2018). To address this need, our work introduces two novel benchmarks in the healthcare domain for domain-incremental continual learning, utilizing real datasets.

Text-to-Image Stable Diffusion. Conditional image generation has been a prominent research area, mainly concerning contributions from Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), Variational Autoencoders (VAE)(Kingma & Welling, 2022), and more recently, diffusion models (Ho et al., 2020; Dhariwal & Nichol, 2021). Stable Diffusion is a text-to-image latent diffusion model (LDM) (Rombach et al., 2021) conditioned on text embeddings of a CLIP text encoder. A diffusion model first learns a forward pass by iteratively introducing noise to the initial image. Subsequently, a backward pass eliminates the noise, recovering the final, generated image. To incorporate text conditions, a cross-attention mechanism is integrated into the U-Net architecture. This allows the infusion of textual information during the image generation process. We build upon the foundation of these models, as discussed later in Section 3. In the context of continual learning, the concurrent work by Smith et al. (2023a) examines the phenomenon of forgetting in the stable diffusion model when new concepts are incrementally introduced. In contrast, our approach involves using diffusion models to generate samples and does not involve sequential fine-tuning of the stable diffusion model.