An Empirical Analysis of Continual Learning for Heterogeneous Medical Visual Question Answering

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Deploying medical visual question answering (MedVQA) systems in real-world clinical settings requires models that adapt to new clinical tasks without forgetting previously acquired knowledge. Continual learning (CL) provides a practical framework for this setting. Despite rapid progress in medical vision-language models, the behavior of CL methods when training these models across heterogeneous MedVQA tasks remains underexplored. This work presents a systematic evaluation of CL for MedVQA across diverse clinical objectives, including classification, multi-label classification, detection, cell counting, and report generation. Specifically, we explore (1) the ability of existing CL methods to mitigate catastrophic forgetting; (2) their sensitivity to task ordering, analyzing how different task sequences influence performance retention and forgetting; and (3) the evolution of low-rank adaptation parameters as new tasks are learned, revealing patterns of weight drift under different CL methods. Our findings suggest that existing CL methods struggle to maintain stability-plasticity balance when tasks with different objectives and supervision formats are interleaved. Code and full experimental setup will be publicly available.

Keywords: Continual Learning · Catastrophic Forgetting · Medical Visual Question Answering · Vision-Language Model.

1 Introduction

Medical visual question answering (MedVQA) has emerged as a powerful paradigm for unified reasoning over medical images and natural language, enabling a wide range of clinical tasks such as disease classification, abnormality detection, and report generation [14,16,23]. Recent advances in medical vision-language models (VLMs), together with parameter-efficient fine-tuning methods [8] such as low-rank adaptation (LoRA) [9], have substantially improved performance across diverse medical datasets, positioning MedVQA systems as promising general-purpose clinical assistants [11]. Most MedVQA systems rely on joint multi-task training, assuming simultaneous access to all task data. However, in real-world clinical settings, data typically arrive incrementally and tasks evolve over time [4,10], requiring models to adapt continuously without forgetting prior knowledge.

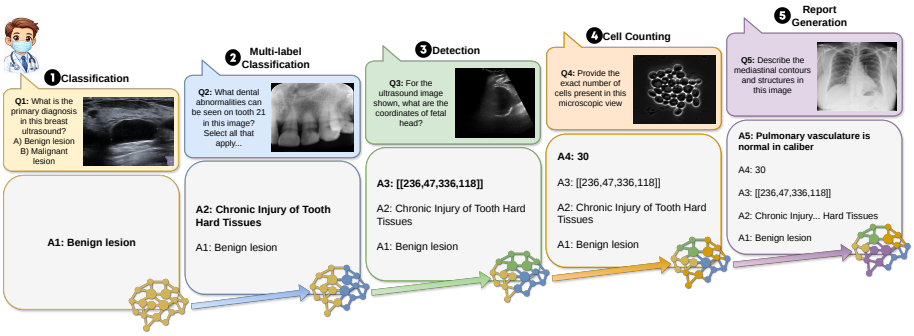

Fig. 1: CL for heterogeneous MedVQA tasks. An illustration of the desired stability-plasticity behavior, where the model is progressively updated to incorporate new tasks while preserving previously acquired knowledge.

Continual learning (CL) [5,3,25,20] provides a principled framework for this setting by enabling sequential task adaptation while achieving an effective *stability-plasticity* trade-off, where the model remains sufficiently *plastic* to adapt to new tasks while maintaining *stability* to preserve previously acquired knowledge and prevent catastrophic forgetting [6]. Although CL has been explored in medical imaging and MedVQA, prior works predominantly focus on domain-incremental settings, where tasks differ in data source or imaging modality but share the same objective and supervision format [4,19,10]. In contrast, real-world clinical workflows require medical VLMs to operate across structurally heterogeneous tasks that differ in objective type, supervision granularity, and semantic constraints on the generated outputs (Fig. 1). This heterogeneity introduces distinct optimization dynamics and cross-task interference patterns that are not captured by conventional CL benchmarks, leaving a critical gap in understanding CL behavior in MedVQA. This motivates three key research questions (RQs): **RQ1.** *Do current CL methods mitigate catastrophic forgetting in heterogeneous MedVQA?*, **RQ2.** *How does task ordering influence performance and forgetting across different CL methods?* and **RQ3.** *How does drift in LoRA weights relate to forgetting behavior across CL methods?*

To answer these questions, we conduct a systematic study of a representative set of CL strategies under sequential training across diverse MedVQA tasks, including classification, multi-label classification, detection, cell counting, and report generation. For **RQ1**, we monitor task-wise performance throughout sequential training and compute forgetting ratios that capture relative performance drops following the learning of new tasks. For **RQ2**, we evaluate the sensitivity of each CL method across three different sequences of tasks with varying complexity based on output structure, supervision granularity, and reasoning demands. For **RQ3**, we visualize sequential LoRA parameter drift and quantify the degree of angular spread, relating these weight dynamics to observed forgetting patterns.

2 Methods

We study MedVQA under a CL setting, where a pre-trained medical VLM is exposed to a sequence of T tasks $\{\mathcal{D}_t\}_{t=1}^{T}$ ($\mathcal{D}_t = \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{N_t}$ denotes the dataset for task t). The tasks are structurally heterogeneous and differ in objective and supervision format, spanning classification, multi-label classification, detection, cell counting, and report generation. Training proceeds sequentially with the objective of optimizing performance on the current task while minimizing performance degradation on previously learned tasks. Importantly, task identity is not provided at inference time; a single shared model must generalize across all tasks without explicit task indicators. The backbone VLM remains frozen throughout training, and only LoRA parameters are updated, placing the stability-plasticity trade-off entirely within the adapter space. The overall goal is to obtain a model that performs strongly across all heterogeneous tasks while effectively mitigating catastrophic forgetting.

To systematically study this setting, we adapt a representative set of CL strategies that cover rehearsal-based, parameter-regularization, knowledge-transfer, and subspace-isolation techniques, reformulating them so that their constraints and updates operate solely on LoRA parameters. We further include two reference baselines: multi-task learning (MTL), which jointly trains a shared LoRA adapter across all tasks as an upper-bound reference, and sequential fine-tuning (SEQ-FT), which updates the same adapter sequentially without forgetting mitigation. The following presents a description of the CL methods:

Replay mitigates catastrophic forgetting by rehearsing samples from previously learned tasks during training on new tasks [2]. We adopt a rehearsal-based variant with a balanced replay buffer that stores samples from prior tasks and mixes them with data from the current task during training [17].

Elastic Weight Consolidation (EWC) [12] regularizes important parameters for previous tasks. We adapt it by restricting consolidation to LoRA parameters. After training task t, parameter importance is estimated using the Fisher information diagonal $F_i^{(t)} \approx \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\left[\left(\frac{\partial \mathcal{L}(x,y;\theta)}{\partial \theta_i}\right)^2\right]_{\theta=\theta_t^*}$, where θ_t^* denotes the convergent parameters after task t. During training on task $t+1$, deviations from θ_t^* are penalized via $\mathcal{L}_{\text{EWC}} = \frac{\lambda_{\text{EWC}}}{2}\sum_i F_i^{(t)}(\theta_i - \theta_{t,i}^*)^2$, yielding the objective $\mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{EWC}}$, where λ_{EWC} controls the stability-plasticity trade-off. We implement multi-task EWC, storing a separate Fisher matrix and parameter snapshot for each completed task and summing their corresponding penalties during subsequent training.

Knowledge Distillation (KD)-based Approach transfers knowledge across tasks using a previously trained model as a reference [15,21,1]. We adapt recent KD-LoRA formulations to a CL setting by using output-level distillation as a regularizer to preserve previously acquired knowledge. In task t, the student model is equipped with LoRA modules and trained on the current dataset $\mathcal{D}_t$, while the teacher model corresponds to the model obtained after completing task $t-1$ and is kept frozen. Given an input $x \sim \mathcal{D}_t$, we encourage the student

to match the teacher's predictive distribution by minimizing an output-level distillation loss $\mathcal{L}_{\mathrm{KD}} = \mathrm{KL}\Big(\sigma\Big(\frac{z(x;\theta^{(t-1)})}{\tau}\Big) \Big\| \sigma\Big(\frac{z(x;\theta^{(t)})}{\tau}\Big)\Big)$, where $z(x;\theta^{(t)})$ and $z(x;\theta^{(t-1)})$ denote the logits produced by the student and teacher models, respectively, $\sigma(\cdot)$ is the softmax function, $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback–Leibler divergence and τ is the distillation temperature. The overall training objective is given by $\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}}$, where $\mathcal{L}_{\mathrm{task}}$ is the task-specific loss on $\mathcal{D}_t$ and λ_{KD} controls the stability-plasticity trade-off.

Soft-masking of Parameter-level Gradient flow (SPG) [13] softly constrains updates to parameters that are important for previous tasks. We adapt it by restricting gradient modulation to LoRA parameters. Parameter importance is estimated via gradient sensitivity and is used to modulate future updates without hard freezing. Let θ denote the trainable parameters. During task t, raw gradients $g_t(\theta) = \nabla_\theta \mathcal{L}_t(\theta)$ are scaled using an accumulated importance mask as $\tilde{g}_t(\theta) = g_t(\theta) \odot \big(1 - a^{(t)}_{\max}(\theta)\big)$, where $a^{(t)}_{\max}(\theta) \in [0,1]$ encodes parameter importance from previous tasks, protecting the parameters of high-importance while keeping others plastic. After training task t, importance is computed by probing the model with data from all observed tasks. For each $k \le t$, gradients are accumulated with respect to θ to obtain the gradient tensor $G_{t,k}(\theta)$. Each gradient tensor is standardized via $S(H) = \tanh\Big(\frac{H-\mu(H)}{\sigma(H)+\varepsilon}\Big)$, where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation. The task-level mask is defined as $m^{(t)}(\theta) = \max_{k\le t} |S(G_{t,k}(\theta))|$, and the accumulated mask used for subsequent training is $a^{(t)}_{\max}(\theta) = \max_{j<t} m^{(j)}(\theta)$, which attenuates updates to previously important parameters while allowing adaptation to new tasks.

Orthogonal Low-Rank Adaptation (O-LoRA) [26] assigns each task its own low-rank adapter and constrains new updates to be orthogonal to previously learned subspaces. For each task t, O-LoRA introduces a new adapter (A_t, B_t) while freezing prior adapters $\{(A_i, B_i)\}_{i<t}$. Let A_t denote the current LoRA projection matrix and $\{A_i\}_{i<t}$ those of earlier tasks. The overlap between A_i and A_t subspaces is penalized using the orthogonality loss $\mathcal{L}_{\mathrm{ortho}} = \sum_{i<t} \|A_i A_t^\top\|_F^2$. The objective for task t becomes $\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{ortho}}\mathcal{L}_{\mathrm{ortho}}$, where λ_{ortho} controls the strength of the constraint. During training, the forward pass composes all previously learned (frozen) adapters with the current adapter, while gradients update only the current one. At inference time, adapters are composed to aggregate task knowledge without modifying the frozen backbone.

3 Experimental Setup

3.1 Dataset and Evaluation Metrics

We conduct experiments on the FLARE-MLLM-2D benchmark [7], which contains image-question-answer triplets collected from 19 publicly available medical datasets. The benchmark covers eight imaging modalities and five clinical tasks, providing a comprehensive and well-suited testbed for evaluating CL under realistic task diversity. The modality distribution across splits is shown in Fig. 2.

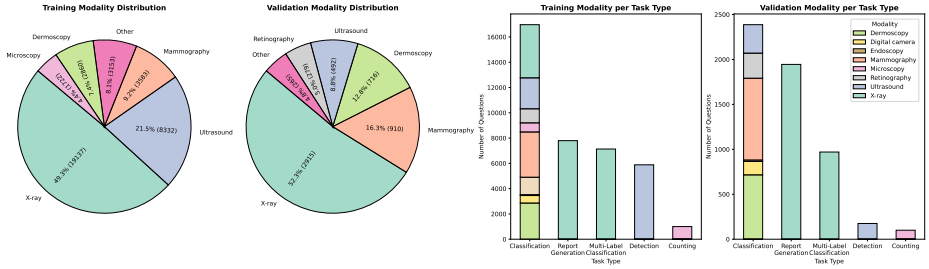

Fig. 2: Distribution of imaging modalities for training and validation sets.

Each task is assessed with a task-specific metric. Classification uses *Balanced Accuracy* to account for class imbalance. Multi-label classification is evaluated with the *Micro-averaged F1 Score*. Detection is measured by *F1 Score* with an IoU threshold of 0.5. Cell counting uses *Mean Absolute Error (MAE)*. Report generation is evaluated with the *GREEN score*, which measures factual correctness and clinical relevance by capturing lexical overlap, semantic consistency, and clinical accuracy [18]. To quantify forgetting, we report the *Forgetting Ratio (FR)*, defined as the relative performance drop for each task between its best performance achieved at any point during sequential training and its performance after the model has finished training on all tasks. For tasks where higher metric values are better, $\mathrm{FR}(t) = \frac{\max_{t \le i \le T}(p_t^{(i)}) - p_t^{(T)}}{\max_{t \le i \le T}(p_t^{(i)})}$. For tasks where lower metric values are better, $\mathrm{FR}(t) = \frac{p_t^{(T)} - \min_{t \le i \le T}(p_t^{(i)})}{\min_{t \le i \le T}(p_t^{(i)})}$.

3.2 Implementation Details

We use `MedGemma-4b-it` [22] as it is pre-trained and instruction-tuned on diverse clinical tasks and imaging modalities, making it well-suited for comprehensive MedVQA evaluation. The model is loaded in 4-bit NF4 quantized format with bfloat16 precision. We apply LoRA to all linear layers with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. Optimization uses AdamW with a learning rate of 2×10^{-4} and linear scheduling. Each experience is trained for 3 epochs with a batch size of 4 and gradient accumulation of 4 steps. The hyperparameter selection is based on the best-performing model reported on the FLARE-MLLM-2D benchmark [7,24]. We use a fixed replay buffer of 100 samples per task, capped by the number of available training samples for low-resource tasks. Following prior literature [1,12,26], we adopt standard hyperparameter settings for the rest of the CL methods: $\lambda_{\mathrm{EWC}} = 5000$, $\lambda_{\mathrm{KD}} = 0.5$, $\tau = 2$, and $\lambda_{\mathrm{ortho}} = 0.5$. We evaluate three curriculum strategies: **Order A** (Easy → Hard): classification, multi-label classification, detection, cell counting, report generation; **Order B** (Hard → Easy): report generation, cell counting, detection, multi-label classification, classification; and **Order C** (Mixed): classification, multi-label classification, report generation, detection, cell counting.

Table 1: Performance comparison of non-CL and CL methods. ↑ denotes higher is better, while ↓ denotes lower is better. FR $\in [0, 1]$ denotes the forgetting ratio (0 = no forgetting, 1 = complete forgetting). FR is not applicable for MTL and for the last task in each order, and is marked as "–". Each metric reflects average task performance across all sequential training stages.

		Classification		Multi-label Classification		Detection		Cell Counting		Report Generation	
Method	Order	Balanced Acc. ↑	FR ↓	F1 Score ↑	FR ↓	F1 Score ↑	FR ↓	MAE ↓	FR ↓	GREEN Score ↑	FR ↓
MTL (Qwen2.5-VL-7B) [24]	–	0.36	–	0.36	–	0.64	–	243.60	–	0.76	–
MTL (InternVL3-8B) [24]	–	0.52	–	0.46	–	0.37	–	301.43	–	0.75	–
MTL (MedGemma-4B)	–	0.61	–	0.55	–	0.16	–	261.26	–	0.72	–
SeqFT	A	0.16	0.94	0.17	0.88	0.04	1.00	371.96	0.34	0.72	–
Replay	A	0.35	0.50	0.42	0.35	0.05	1.00	360.84	0.21	0.73	–
	B	0.57	–	0.46	0.38	0.04	0.95	361.56	0.08	0.62	0.19
	C	0.35	0.47	0.41	0.36	0.08	1.00	323.12	–	0.76	0.00
EWC	A	0.48	0.40	0.45	0.47	0.13	0.24	317.72	0.00	0.72	–
	B	0.52	–	0.40	0.34	0.06	0.83	263.43	0.02	0.71	0.01
	C	0.46	0.20	0.38	0.35	0.07	0.67	281.99	–	0.73	0.00
KD	A	0.55	0.30	0.41	0.38	0.00	0.00	361.88	0.00	0.70	–
	B	0.43	–	0.39	0.00	0.00	1.00	389.42	0.10	0.64	0.33
	C	0.48	0.20	0.36	0.25	0.00	0.00	320.73	–	0.65	0.13
SPG	A	0.39	0.85	0.41	0.90	0.03	1.00	313.67	0.07	0.71	–
	B	0.49	–	0.51	0.09	0.08	0.68	323.96	0.00	0.34	0.90
	C	0.42	0.31	0.34	0.29	0.09	0.64	232.92	–	0.71	0.00
O-LoRA	A	0.38	1.00	0.34	1.00	0.03	1.00	403.79	0.27	0.71	–
	B	0.44	–	0.31	0.82	0.03	1.00	388.45	0.08	0.30	0.99
	C	0.25	0.95	0.23	1.00	0.03	0.57	408.00	–	0.26	1.00

4 Results and Discussion

We conduct a systematic evaluation to assess the performance of existing CL methods when a medical VLM is trained sequentially on heterogeneous MedVQA tasks. Our analysis explores three key aspects: (i) the extent of catastrophic forgetting and the stability-plasticity trade-off, (ii) the sensitivity to task ordering, and (iii) the weight drift behavior of LoRA adapters. We organize our experimental analysis around three RQs.

RQ1. Do current CL methods mitigate catastrophic forgetting in heterogeneous MedVQA? Results in Table 1 show that current CL methods mitigate but do not eliminate catastrophic forgetting under heterogeneous multimodal supervision. SEQ-FT leads to severe degradation across all tasks, confirming strong cross-task interference. Among CL approaches, EWC demonstrates the most consistent stability-plasticity balance across task types and orders. While KD occasionally achieves comparable or lower forgetting on specific tasks, it shows reduced plasticity on newly introduced objectives, such as counting and report generation, in certain orders. Replay improves over naïve fine-tuning but does not consistently control forgetting. SPG shows order-dependent instability, and O-LoRA exhibits larger forgetting spikes. Performance in detection remains consistently low across all configurations. This trend is consistent with prior studies, which report limited detection performance even when varying backbone architectures or fine-tuning task-specific LoRA [7,24].

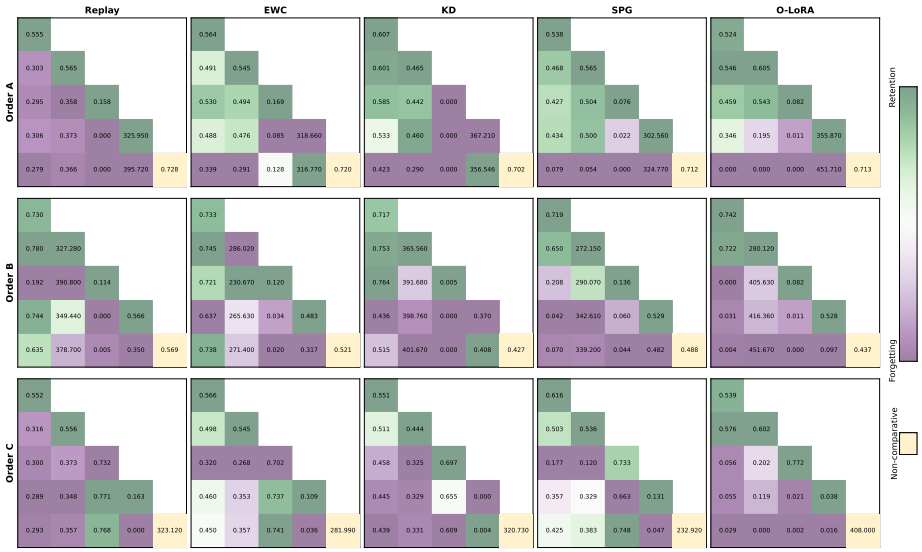

Fig. 3: Performance and forgetting trends across CL methods. Colors encode relative performance change across training stages.

RQ2. How does task ordering influence performance and forgetting across different CL methods? The triangular matrices in Fig. 3, where each entry(i, j) represents performance on Task$_j$ after training on Task$_i$, show that task ordering significantly affects both final performance and forgetting behavior, and sensitivity varies across methods. KD and EWC are comparatively more stable to permutation, exhibiting smaller performance fluctuations across curricula. O-LoRA is highly sensitive to ordering, while Replay and SPG show intermediate sensitivity. We observed that a central factor driving order effects is report generation. As a long-form autoregressive objective, it shifts the model toward producing extended free-form text rather than concise, format-constrained outputs. After training on report generation, most methods generate longer responses even for tasks requiring short answers, altering previously learned output formats, and increasing forgetting when introduced early. This behavior is less evident in KD, as output distillation constrains decoding shifts.

RQ3. How does drift in LoRA weights relate to forgetting behavior across CL methods? In Fig. 4, we visualize the angles between the LoRA directions, measuring similarity via the Frobenius inner product $\mathrm{sim}(\mathbf{t_1}, \mathbf{t_2}) = \frac{\langle \mathbf{t_1}, \mathbf{t_2} \rangle_F}{\|\mathbf{t_1}\|_F \, \|\mathbf{t_2}\|_F}$, where $\langle \mathbf{t_1}, \mathbf{t_2} \rangle_F = \mathrm{Tr}(\mathbf{t}_1 \mathbf{t}_2^\top)$. The figure reveals distinct LoRA parameter shift patterns across CL methods that align with their observed forgetting behavior. Replay exhibits a broad angular spread, indicating substantial directional drift as new tasks are introduced. In contrast, EWC and KD maintain tighter distributions, suggesting closer alignment with earlier tasks, thus improving stability. SPG shows moderate dispersion. O-LoRA exhibits progressive

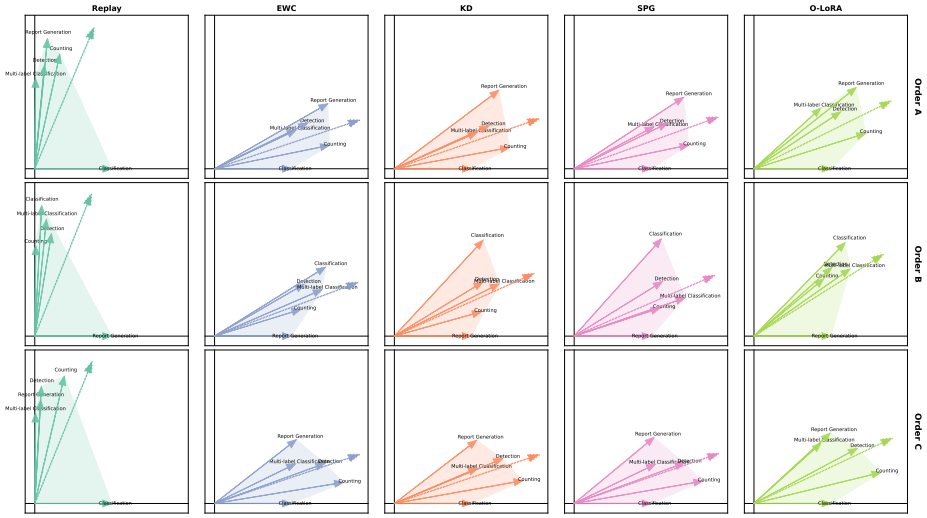

Fig. 4: LoRA parameter directions across sequential tasks. Angles are computed in the high-dimensional LoRA space using Frobenius inner-product similarity between $t-1$ and t and visualized as 2D arrows via polar mapping. The dotted arrow shows the mean drift direction.

directional shift across tasks, indicating cumulative drift. When report generation is introduced early (Orders B and C), the initial angular deviation increases, and the overall drift is amplified. Overall, these results indicate that CL methods differ in reshaping the underlying low-rank parameter space in ways that correlate with the observed forgetting behavior.

5 Conclusion

This study presents a comprehensive evaluation of CL for heterogeneous MedVQA, including classification, multi-label classification, detection, cell counting, and report generation. Our results show that existing CL strategies struggle to mitigate catastrophic forgetting and maintain a stability-plasticity balance when tasks differ in objectives and supervision formats. Performance retention varies significantly across methods and is highly sensitive to task order, indicating limited robustness in realistic sequential training. Analysis of LoRA dynamics further shows that some methods maintain closer alignment of parameter updates across tasks, whereas others induce larger directional shifts, reflecting varying degrees of cross-task interference. These findings show that parameter-level constraints alone are insufficient; robust CL for heterogeneous MedVQA also requires explicit functional alignment to preserve output behavior across arbitrary task sequences. Future work may investigate more advanced methods within the most promising CL strategies, as well as evaluate larger VLMs and 3D medical imaging tasks.

References

1. Azimi, R., Rishav, R., Teichmann, M., et al.: Kd-lora: A hybrid approach to efficient fine-tuning with lora and knowledge distillation. arXiv preprint arXiv:2410.20777 (2024)
2. Bagus, B., Gepperth, A.: An investigation of replay-based approaches for continual learning. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–9. IEEE (2021)
3. Biesialska, M., Biesialska, K., Costa-Jussa, M.R.: Continual lifelong learning in natural language processing: A survey. In: Proceedings of the 28th international conference on computational linguistics. pp. 6523–6541 (2020)
4. Bruno, P., Quarta, A., Calimeri, F.: Continual learning in medicine: A systematic literature review. Neural Processing Letters **57**(1), 2 (2025)
5. Chen, Z., Liu, B.: Lifelong machine learning. Morgan & Claypool Publishers (2018)
6. Chen, Z., Liu, B.: Continual learning and catastrophic forgetting. In: Lifelong Machine Learning, pp. 55–75. Springer (2022)
7. FLARE-MedFM: FLARE-Task5-MLLM-2D medical multimodal dataset. https://huggingface.co/datasets/FLARE-MedFM/FLARE-Task5-MLLM-2D (2025), accessed: 2026
8. Han, Z., Gao, C., Liu, J., Zhang, J., Zhang, S.Q.: Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv:2403.14608 (2024)
9. Hu, E.J., Shen, Y., Wallis, P., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
10. Jiang, N., Zheng, W., Gu, X., et al.: Omnidoctor: Towards llm-centric lifelong learning for new emerging medical vqa tasks. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 6567–6575 (2025)
11. Kalpélbé, B.C., Adaambiik, A.G., Peng, W.: Vision language models in medicine. arXiv preprint arXiv:2503.01863 (2025)
12. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017). https://doi.org/10.1073/pnas.1611835114
13. Konishi, T., Kurokawa, M., Ono, C., et al.: Parameter-level soft-masking for continual learning. In: Proceedings of the 40th International Conference on Machine Learning (2023)
14. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 180251 (2018)
15. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
16. Lin, Z., Zhang, D., Tao, Q., et al.: Medical visual question answering: A survey. Artificial Intelligence in Medicine **143**, 102611 (2023)
17. Merlin, G., Lomonaco, V., Cossu, A., et al.: Practical recommendations for replay-based continual learning methods. In: Image Analysis and Processing – ICIAP 2022 Workshops. pp. 548–559. Springer (2022)
18. Ostmeier, S., Xu, J., Chen, Z., et al.: Green: Generative radiology report evaluation and error notation. In: Findings of the Association for Computational Linguistics: EMNLP 2024. p. 374–390. Association for Computational Linguistics (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.21, http://dx.doi.org/10.18653/v1/2024.findings-emnlp.21

19. Qazi, M.A., Hashmi, A.U.R., Sanjeev, S., et al.: Continual learning in medical imaging: a survey and practical analysis. ACM Computing Surveys (2024)
20. Qu, H., Rahmani, H., Xu, L., et al.: Recent advances of continual learning in computer vision: An overview. IET Computer Vision **19**(1), e70013 (2025)
21. Rebuffi, S.A., Kolesnikov, A., Sperl, G., et al.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
22. Sellergren, A., Kazemzadeh, S., Jaroensri, T., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
23. Shaaban, M.A., Saleem, T.J., Papineni, V.R.K., et al.: Motor: Multimodal optimal transport via grounded retrieval in medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 459–469. Springer (2025)
24. Shaaban, M.A., Saqib, A., Hardan, S.E., Taratynova, D., Saleem, T.J., Yaqub, M.: Me-vlip: A modular and efficient vision-language framework for generalizable medical image parsing. In: MICCAI 2025 FLARE Challenge
25. Wang, L., Zhang, X., Su, H., et al.: A comprehensive survey of continual learning: Theory, method and application. IEEE transactions on pattern analysis and machine intelligence **46**(8), 5362–5383 (2024)
26. Wang, X., Chen, T., Ge, Q., et al.: Orthogonal subspace learning for language model continual learning. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 10658–10671 (2023)