Replay-free Sequential Fine-tuning of Medical VLMs

He Li LIHE22@MAILS.TSINGHUA.EDU.CN
Tsinghua University, People's Republic of China

Yuhui Zhang YUHUIZ@STANFORD.EDU
Xiaohan Wang XHANWANG@STANFORD.EDU
Serena Yeung-Levy SYYEUNG@STANFORD.EDU
Stanford University, United States

Abstract

Catastrophic forgetting severely limits the adaptation of Vision Language Models (VLMs) for medical applications, where sequential learning on private data is often necessary. We propose that this issue stems from insufficient regularization and demonstrate that regularizing parameter updates during fine-tuning effectively mitigates forgetting without harming new task performance. To validate our method in a clinical context, we introduce **Medical-CL**, a new continual learning benchmark spanning pathology, cell microscopy, radiology, and surgery. Our streamlined, replay-free approach proves highly effective on this benchmark, offering a practical path toward building comprehensive, continually-learning medical VLMs and advancing the development of medical AI.

Keywords: Vision Language Models, Continual Learning, Catastrophic Forgetting.

Data and Code Availability We use publicly available datasets: MLLM-CL Benchmark, BSCCM, PitVis-2023, PathVQA, ROCOv2. We will make code and data publicly available.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

The emergence of Vision Language Models (VLMs) represents a significant milestone in artificial intelligence (Alayrac et al., 2022; Liu et al., 2023; Achiam et al., 2023). Building upon this foundational success, the field has increasingly focused on fine-tuning these powerful models for specialized medical applications, aiming to create robust medical VLMs.

However, the catastrophic forgetting phenomenon (Zhai et al., 2024; Shuttleworth et al., 2024) is an unavoidable issue for the development of medical VLM. This issue manifests as a severe degradation in a model's performance on previously learned tasks after it has been fine-tuned for a new specialization (McCloskey and Cohen, 1989). Within the context of medical VLMs, fine-tuning on a specific dataset could inadvertently compromise its general reasoning abilities or diminish its acquired knowledge pertinent to other medical fields. Furthermore, given that a significant volume of medical data is private or sensitive and cannot be easily shared, sequential learning on different datasets is an unavoidable necessity for building comprehensive medical VLMs under current constraints.

Drawing inspiration from prior research into the loss landscapes of Large Language Models (Chen et al., 2025) and the significance of orthogonal subspaces during fine-tuning (Wang et al., 2023), we hypothesize that the forgetting phenomenon arises from insufficient regularization during the adaptation process. Based on reasonable assumptions derived from landscape and subspace theory, we present the rationale for the efficacy of *regularizing parameter updates*.

Our experimental results demonstrate that *regularization applied to parameter updates can effectively mitigate forgetting without compromising performance on the fine-tuning task*. This holds true across multiple sequential fine-tuning scenarios, including domain-continual learning (Zhao et al., 2025). Empirical evaluations reveal that our simple approach surpasses the performance of existing methods, particularly those that depend on a data replay buffer. This outcome highlights the strong, inherent capacity of Vision-Language Models for continual learning.

Subsequently, we extend our investigation to various medical domains, curating a novel medical continual learning benchmark, which we term **Medical-**

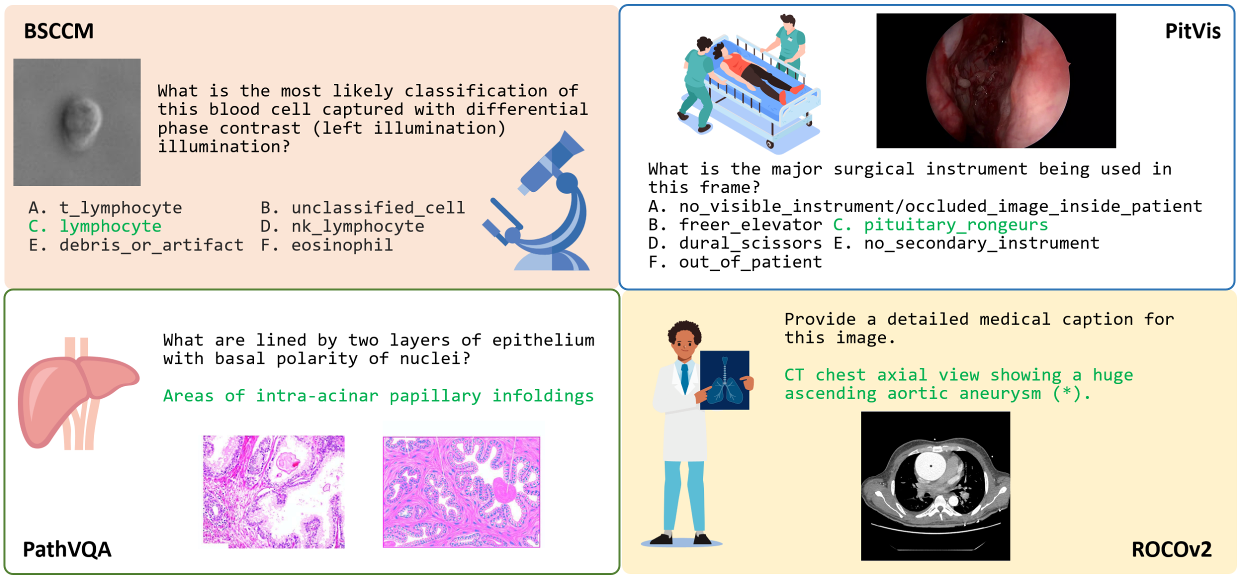

Figure 1: An Illustration of the Proposed Medical-CL Benchmark

CL, from a collection of public datasets (Pinkard et al., 2024; Das et al., 2025; He et al., 2020; Rückert et al., 2024). The **Medical-CL** benchmark incorporates a variety of question formats, including åmultiple-choice, short-answer, and image captioning, to ensure a comprehensive evaluation. Adhering to a protocol similar to the **MLLM-CL** continual learning benchmark (Zhao et al., 2025), we confirm that our proposed method remains highly effective on this new benchmark.

In summary, this paper contributes a comprehensive experimental investigation and corresponding theoretical analysis of the catastrophic forgetting phenomenon during VLM fine-tuning. Our work offers practical guidance for adapting these models to medical domains while preserving their previous capabilities. This facilitates the sequential fine-tuning of VLMs on new medical data with minimal performance loss on previously learned tasks. We anticipate that this research will benefit practitioners in the field and advance the development of more robust and theoretically grounded VLMs. Code, checkpoints and data are available through annoymous links in Appendix A.

2. Challenge: Sequentially Fine-tuning of VLMs

2.1. MLLM-CL Benchmark for General Continual Learning

The challenge of general sequential fine-tuning is analogous to the continual learning problem. Accordingly, we adopt the MLLM-CL benchmark and its evaluation protocols as proposed in a recent study (Zhao et al., 2025). This sequential learning benchmark comprises five distinct domains: Remote Sensing (**RSVQA**), Medicine (**PathVQA**), Autonomous Driving (**DriveLM**), Science (**AI2D, SciVerse, MapQA, TQA**), and Finance (**StockQA**).

For simplicity, these tasks are denoted as **RS, Med, AD, Sci, and Fin**, respectively. In our experiments, we follow the sequential fine-tuning order established in the original MLLM-CL study: **RS** → **Med** → **AD** → **Sci** → **Fin**. Training Details are provided in Appendix C.

2.2. Evaluation Metrics

To ensure a fair and direct comparison, our evaluation protocol strictly adheres to the methodology outlined in MLLM-CL (Zhao et al., 2025). We report two primary metrics: *Last* and *Average*. The *Last* metric represents the average accuracy across all previously

Method	**Last**					**Average**				
	RS (%)	**Med (%)**	**AD (%)**	**Sci (%)**	**Fin (%)**	**RS (%)**	**Med (%)**	**AD (%)**	**Sci (%)**	**Fin (%)**
Zero-shot	32.29	28.28	15.59	35.55	62.56	-	-	-	-	-
w/ replay buffer										
LoRA	29.57	29.19	7.09	19.55	63.60	<u>80.87</u>	58.60	38.95	36.41	36.78
MoELoRA	40.23	23.58	5.19	18.35	74.89	80.00	56.91	34.69	31.70	31.36
O-LoRA	76.21	51.34	36.50	42.64	90.20	80.13	70.23	61.35	53.34	59.38
L2P	75.21	38.50	32.31	41.05	88.05	80.09	68.64	54.79	48.68	55.02
ModalPrompt	64.77	38.60	20.61	29.98	88.22	80.11	60.99	50.67	41.97	48.44
HiDe-LLaVA	75.36	39.23	37.17	45.02	81.89	**81.51**	62.37	49.37	50.61	55.73
MR-LoRA	**79.87**	**62.71**	<u>51.89</u>	**52.48**	89.69	80.82	**72.19**	<u>65.41</u>	**62.52**	**67.31**
IncLoRA (Ours)	77.43	<u>62.57</u>	**52.00**	**52.48**	<u>90.41</u>	78.30	71.93	65.38	62.12	66.98
SeqFull (Ours)	<u>78.94</u>	62.45	51.50	<u>52.08</u>	**91.21**	75.62	<u>72.16</u>	**65.77**	<u>62.32</u>	<u>67.24</u>
w/o replay buffer										
LoRA	26.75	25.76	0.79	18.69	70.44	<u>80.72</u>	59.68	40.51	18.64	28.49
MoELoRA	21.42	25.29	0.79	17.01	60.34	80.05	57.26	37.03	19.65	24.97
O-LoRA	62.68	35.17	16.93	34.44	92.16	80.22	67.56	51.51	44.28	48.28
L2P	63.82	34.63	22.96	38.58	**92.98**	80.02	68.86	51.57	45.12	50.59
ModalPrompt	65.99	37.35	23.27	37.61	87.60	80.11	59.66	46.86	42.97	50.36
HiDe-LLaVA	41.17	30.33	18.73	37.08	<u>92.21</u>	**80.91**	65.47	39.78	32.92	43.90
IncLoRA (Ours)	<u>77.20</u>	<u>58.97</u>	<u>51.43</u>	<u>47.44</u>	90.24	77.59	<u>71.59</u>	<u>64.40</u>	<u>60.22</u>	<u>65.06</u>
SeqFull (Ours)	**79.10**	**61.22**	**52.36**	**50.52**	91.29	77.06	**72.75**	**66.09**	**62.49**	**67.44**

Table 1: Results for domain continual learning in MLLM-CL benchmark. We highlight **the best result** and <u>the second best result</u> separately for *w/ replay buffer* and *w/o replay buffer*.

learned tasks after the model has completed training on the final task in the sequence. The *Average* metric captures performance throughout the entire training process, defined as the mean of the average accuracies calculated after each sequential task is learned.

2.3. Medical-CL Benchmark for Medical Continual Learning

To specifically evaluate sequential fine-tuning within the medical domain, we introduce a novel benchmark, termed **Medical-CL**. This benchmark is curated from several publicly available datasets, including BSCCM (Pinkard et al., 2024), PitVis (Das et al., 2025), PathVQA (He et al., 2020), and ROCOv2 (Rückert et al., 2024).

These tasks are denoted as **Cell, Sur, Path, and Rad**, respectively. We follow MLLM-CL and use a randomized order of **Path** → **Sur** → **Cell** → **Rad**. For **Cell, Sur and Path**, we use answer accuracy as the metric, while for **Rad**, we adapt BLEU (Papineni et al., 2002) and rescale it to 0-100 to fit the regular percentage accuracy. Training Details are provided in Appendix D.

As illustrated in Figure 1, each dataset was adapted for VLM evaluation. Our benchmark comprises four distinct medical datasets, spanning a range of scenarios from cellular-level analysis and surgical procedures to high-level pathology and radiology interpretation. This diversity provides a robust framework for evaluating sequential learning capabilities in specialized medical contexts.

3. Analysis: Why the Forgetting is Happening?

Research indicates that pretrained Large Language Models (LLMs) have loss landscapes with wide, flat, and anisotropic basins, where model performance is stable (Chen et al., 2025; Xu et al., 2024). The pretraining on web-scale data creates a general foundational basin. The process of fine-tuning can be seen as optimizing a more specialized sub-basin within this foundational one to adapt the model to a specific target task.

For a pretrained model f_θ and a task $\mathcal{T}$, the loss landscape $\mathcal{L}_{f_\theta,\mathcal{T}}$ contains a task-specific sub-basin. Catastrophic forgetting occurs when subsequent parameter updates push the model outside of this specialized sub-basin. To analyze this, we can decompose the parameter space into two orthogonal subspaces: a **robust subspace** $\mathcal{R}_{f_\theta,\mathcal{T}}$, characterized by low loss curvature where parameter changes have

IncLoRA	Path	Sur	Cell	Rad	Average
Zero-shot	36.01	31.45	4.99	12.86	21.11
Path	68.85				68.84
Sur	63.83	53.43			58.63
Cell	59.93	51.62	84.34		65.30
Rad	53.58	47.12	75.58	18.12	48.57

Table 2: Performance of **IncLoRA** on the Medical-CL benchmark without a replay buffer.

SeqFull	Path	Sur	Cell	Rad	Average
Zero-shot	36.01	31.45	4.99	12.86	21.11
Path	61.17				61.17
Sur	59.85	58.03			58.94
Cell	59.24	56.47	82.09		65.94
Rad	54.54	57.54	80.27	16.20	52.09

Table 3: Performance of **SeqFull** on the Medical-CL benchmark without a replay buffer.

minimal impact, and a **sensitive subspace** $\mathcal{S}_{f_\theta,\mathcal{T}}$, defined by high loss curvature where performance is acutely sensitive to changes.

The key to preventing catastrophic forgetting is to regularize parameter updates to avoid disrupting the sensitive subspace $\mathcal{S}_{f_\theta,\mathcal{T}}$ of previously learned tasks. Common fine-tuning strategies achieve this through two primary forms of regularization:

- **Low Learning Rates** act as a *soft regularization*. By discouraging large steps, this approach biases the optimization process to remain within the robust subspace $\mathcal{R}_{f_\theta,\mathcal{T}}$, reducing the likelihood of venturing into the sensitive subspace.
- **Parameter-Efficient Fine-Tuning (PEFT)** methods like LoRA impose a *structural regularization*. They restrict all updates to a predefined, low-dimensional parameter subspace. This inherently limits the dimensionality of the sensitive subspace that can be altered, thereby preserving knowledge from prior tasks by design.

In essence, both strategies mitigate forgetting by constraining parameter updates, either in magnitude or direction, to protect the sensitive dimensions critical for retaining previously acquired knowledge.

4. Solution: Limiting the Parameter Update

4.1. Ablation Study on General MLLM-CL

As presented in Table 2, we conduct a comprehensive evaluation of our proposed methods, **IncLoRA** and **SeqFull**, on the general-purpose MLLM-CL benchmark. **IncLoRA** will reinitialize a new LoRA for each task and merge the LoRA weight into the model after learning each task. Then, the next task in stream will use this merged model as the new base model and repeat the process. **SeqFull** as it name, will sequentially train the all the LLM Backbone parameters without any trick.

In the setting where a replay buffer (details in Appendix B) is utilized, many contemporary methods employ sophisticated mechanisms to mitigate forgetting. As evidenced by the results, our simple, trick-free methods achieve performance that is highly comparable to the state-of-the-art. For instance, our **SeqFull** method achieves 78.94% on the **RS** task under the **Last** metric, closely trailing the 79.87% of the more complex **MR-LoRA**, while simultaneously outperforming it in the **Fin** domain.

The advantages of our methodology become even more pronounced in the more challenging and realistic scenario without a replay buffer, since medical data involves privacy and any possible leakage from replay buffer is unacceptable. In this setting, **IncLoRA** and **SeqFull** consistently outperforms all other competing methods, establishing new benchmarks across most domains.

4.2. Application on Medical-CL

To evaluate the effectiveness of our proposed methods in a specialized and privacy-sensitive domain, we applied IncLoRA and SeqFull to the Medical-CL benchmark, as shown in Table 2 and Table 3. The results demonstrate strong performance in this challenging, buffer-free setting.

Both methods significantly outperform the zero-shot baseline across all medical subdomains (Pathology, Surgery, Cell, and Radiology). For instance, after being trained on the full sequence of tasks, **SeqFull** achieves a final average score of 52.09, while **IncLoRA** achieves 48.57, both substantial improvements over the 21.11 baseline. Notably, both models show a strong ability to acquire new knowledge, with Cell performance reaching 84.34 for **IncLoRA** and 82.09 for **SeqFull**. These results highlight the viability of our simple yet effective continual learning strategies for specialized applications where data privacy is paramount.

5. Discussion

Our findings have significant implications for medical AI, offering a replay-free continual learning method that addresses data privacy by enabling institutions to fine-tune VLMs on local datasets. Future work could focus on optimizing regularization strategies for diverse medical modalities or assessing long-term performance. These next steps are vital for developing theoretically grounded, adaptable, and scalable VLMs that can safely learn across the vast landscape of medical knowledge.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Huanran Chen, Yinpeng Dong, Zeming Wei, Yao Huang, Yichi Zhang, Hang Su, and Jun Zhu. Understanding pre-training and fine-tuning from loss landscape perspectives. *arXiv preprint arXiv:2505.17646*, 2025.

Adrito Das, Danyal Z Khan, Dimitrios Psychogyios, Yitong Zhang, John G Hanrahan, Francisco Vasconcelos, You Pang, Zhen Chen, Jinlin Wu, Xiaoyang Zou, et al. Pitvis-2023 challenge: Workflow recognition in videos of endoscopic pituitary surgery. *Medical Image Analysis*, page 103716, 2025.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Henry Pinkard, Cherry Liu, Fanice Nyatigo, Daniel A Fletcher, and Laura Waller. The berkeley single cell computational microscopy (bsccm) dataset. *arXiv preprint arXiv:2402.06191*, 2024.

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, et al. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1): 688, 2024.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full finetuning: An illusion of equivalence. *arXiv preprint arXiv:2410.21228*, 2024.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. *arXiv preprint arXiv:2310.14152*, 2023.

Yichu Xu, Xin-Chun Li, Lan Li, and De-Chuan Zhan. Visualizing, rethinking, and mining the loss landscape of deep neural networks. *arXiv preprint arXiv:2405.12493*, 2024.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, pages 202–227. PMLR, 2024.

Hongbo Zhao, Fei Zhu, Rundong Wang, Gaofeng Meng, and Zhaoxiang Zhang. Mllm-cl: Continual learning for multimodal large language models. *arXiv preprint arXiv:2506.05453*, 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

Appendix A. Annoymous Links

We provide fully annoymous links to our supplementary materials here.

Code
https://anonymous.4open.science/r/replay-free-finetuning-medical-vlm-BB37

Model Checkpoints
https://huggingface.co/Replay-Free-Finetuning-Medical-VLM/checkpoints

Medical-CL Dataset
https://huggingface.co/datasets/Replay-Free-Finetuning-Medical-VLM/Medical-CL

Appendix B. Fine-tuning Protocol

Base Model. For all the evaluations in this paper, we adapt `LLaVA-7B` as our base model for fine-tuning. The checkpoint could be downloaded from https://huggingface.co/llava-hf/llava-1.5-7b-hf.

Fine-tuning Settings. For **IncLoRA** fine-tuning, we set the LLM backbone as the LoRA target and unfreeze the projector. For **SeqFull** fine-tuning, we all the paramters of LLM backbone and projector to be trainable.

Training Framework. Our experimental framework is built upon the LLaMA-Factory repository (Zheng et al., 2024). The training configurations adhere to the official guidelines provided in the `LLaVA` model repositories (Liu et al., 2023).

Prompt Templates. For the `LLaVA` model, we utilized the corresponding system prompt templates provided within the LLaMA-Factory framework. For all the evaluations except the image captioning for **Radiology**, we turn all the questions into multiple-choice and add format instruction in prompt to avoid the influence of the mismatch of output format.

Replay Buffer. We exactly follow the setting in **MLLM-CL** Zhao et al. (2025), specifically, for each task, we collect a replay data buffer of size 20 samples. Then, for every downstream sequential fine-tuning, we directly hybrid the all the replay data into the training data. No over-sampling is implemented.

Appendix C. MLLM-CL Fine-tuning Hyperparameters

The training length for every task is aligned to MLLM-CL (Zhao et al., 2025) for fair comparison.

Config	Value
optimizer	AdamW
batch size	64
lr schedule	cosine decay
lr warmup ratio	0.1
base lr	8×10^{-5}
epoch for RS	1
epoch for Med	3
epoch for AD	1
epoch for Sci	2
epoch for Fin	1
LoRA rank	8

Table 4: Hyperparameters of **IncLoRA** in MLLM-CL Benchmark *w/o replay buffer*.

Config	Value
optimizer	AdamW
batch size	64
lr schedule	cosine decay
lr warmup ratio	0.1
base lr	8×10^{-5}
epoch for RS	1
epoch for Med	3
epoch for AD	1
epoch for Sci	2
epoch for Fin	1
LoRA rank	16

Table 5: Hyperparameters of **IncLoRA** in MLLM-CL Benchmark *w/ replay buffer*.

Config	Value
optimizer	AdamW
batch size	16
lr schedule	cosine decay
lr warmup ratio	0.1
base lr	1×10^{-6}
epoch for RS	1
epoch for Med	3
epoch for AD	1
epoch for Sci	2
epoch for Fin	1

Table 6: Hyperparameters of **SeqFull** in MLLM-CL Benchmark *w/o replay buffer*.

Config	Value
optimizer	AdamW
batch size	16
lr schedule	cosine decay
lr warmup ratio	0.1
base lr	1×10^{-6}
epoch for RS	1
epoch for Med	3
epoch for AD	1
epoch for Sci	2
epoch for Fin	1

Table 7: Hyperparameters of **SeqFull** in MLLM-CL Benchmark *w/ replay buffer*.

Appendix D. Medical-CL Fine-tuning Hyperparameters

All the experiment in this part is *w/o replay buffer*.

Config	**Value**
optimizer	AdamW
batch size	64
lr schedule	cosine decay
lr warmup ratio	0.1
base lr	1×10^{-6}
step for Path	2000
step for Sur	2000
step for Cell	2000
step for Rad	2000
LoRA rank	16

Table 8: Hyperparameters of **IncLoRA** in Medical-CL Benchmark.

Config	**Value**
optimizer	AdamW
batch size	16
lr schedule	cosine decay
lr warmup ratio	0.1
base lr	1×10^{-6}
step for Path	2000
step for Sur	2000
step for Cell	2000
step for Rad	2000

Table 9: Hyperparameters of **SeqFull** in Medical-CL Benchmark.