Continual Retinal Vision-Language Pre-training upon Incremental Imaging Modalities

Yuang Yao[1,2], Ruiqi Wu[1,2], Yi Zhou[1,2⋆], and Tao Zhou[3]

[1] School of Computer Science and Engineering, Southeast University, China
[2] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Ministry of Education, China
[3] Nanjing University of Science and Technology, China
{yaoyuang@seu.edu.cn, yizhou.szcn@gmail.com}

Abstract. Traditional fundus image analysis models focus on single-modal tasks, ignoring fundus modality complementarity, which limits their versatility. Recently, retinal foundation models have emerged, but most still remain modality-specific. Integrating multiple fundus imaging modalities into a single foundation model is valuable. However, in dynamic environments, data from different modalities often arrive incrementally, necessitating continual pre-training. To address this, we propose RetCoP, the first continual vision-language pre-training framework in the fundus domain, which incrementally integrates image and text features from different imaging modalities into a single unified foundation model. To mitigate catastrophic forgetting in continual pre-training, we introduce a rehearsal strategy utilizing representative image-text pairs and an off-diagonal information distillation approach. The former allows the model to revisit knowledge from previous stages, while the latter explicitly preserves the alignment between image and text representations. Experiments show that RetCoP outperforms all the compared methods, achieving the best generalization and lowest forgetting rate. **The code can be found at https://github.com/Yuang-Yao/RetCoP**.

Keywords: Continual Pre-training · Vision-Language Pre-training · Multi-modality Fundus Images.

1 Introduction

Fundus imaging reveals eye and systemic diseases, enabling early diagnosis and intervention [3]. Traditional models focus on modality-specific tasks [17,9], requiring separate models and limiting cross-task knowledge sharing. Recently, foundation models in fundus imaging have gained attention [5,15,18]. By pre-training on large-scale datasets, foundation models can provide strong initial weights for downstream tasks, enhancing transfer learning efficiency. However, most current models remain single-modality, hindering generalizability for complex clinical tasks due to insufficient multi-modal information integration.

⋆ Corresponding author: Yi Zhou

Fundus imaging techniques are diverse, such as Color Fundus Photography (CFP), Fluorescein Fundus Angiography (FFA), and Optical Coherence Tomography (OCT), each offering diverse perspectives on retinal structure and disease. Thus, multi-modal foundational pre-training for fundus image analysis holds significant value and potential advancement. In dynamic environments, collecting all modal data at the beginning of pre-training is often impractical, as data from different imaging modalities typically arrives incrementally. Moreover, for new modalities, starting training from scratch is inefficient, necessitating continuous model updates and incremental learning during pre-training.

In continual pre-training, catastrophic forgetting—where new-stage training impairs past knowledge—is a major challenge in dynamic environments [16]. Unlike class-incremental learning, continual pre-training lacks explicit class concepts, requiring the model to learn general representations for each fundus modality and integrate them into a unified space [29]. However, later-stage features often disrupt earlier representation spaces, undermining generalization for previous modalities and worsening forgetting. Thus, efficient continual pre-training algorithms are essential. Existing methods like MedCoSS [27] and CMDK [22] use rehearsal-based self-supervised learning but rely solely on image data, neglecting paired textual descriptions. We insist that text bridges different imaging modalities. If paired image-text data can be fully utilized, it can achieve semantic understanding aligned with images and improve generalization performance. Especially in the medical domain, text descriptions enriched with expert knowledge can further improve pre-training quality. While continual vision-language pre-training has been explored in natural images, such as CTP [29] and IncCLIP [26], these methods focus on class-incremental tasks in natural image domains and are not directly applicable to medical modality-incremental problems. Furthermore, recent CLIP-based continual learning methods (e.g., ZSCL [28]) primarily focus on adapting pretrained models rather than continual pretraining from scratch, making them ineffective for our challenge.

To address these challenges, we propose RetCoP, short for **Ret**inal **Co**ntinual **P**re-training. To the best of our knowledge, ReCoP is the first continual vision-language pre-training framework in the retinal domain. Our framework aligns image-text pairs across incrementally arriving fundus modalities using contrastive pre-training, integrating multi-modal knowledge into a single foundation model. To mitigate catastrophic forgetting in continual training, we introduce a rehearsal strategy based on representative joint embeddings and an off-diagonal information distillation approach. The rehearsal strategy computes image-text joint embeddings using similarity-weighted representations at each stage and selects the most representative image-text pairs via K-means sampling. The off-diagonal information distillation preserves the similarity matrix from the previous stage, maintaining image-text alignment in continual contrastive learning. RetCoP outperforms all comparison methods across downstream tasks, demonstrating superior generalization and minimal forgetting across all three fundus modalities at different training stages. Our main contributions are as follows:

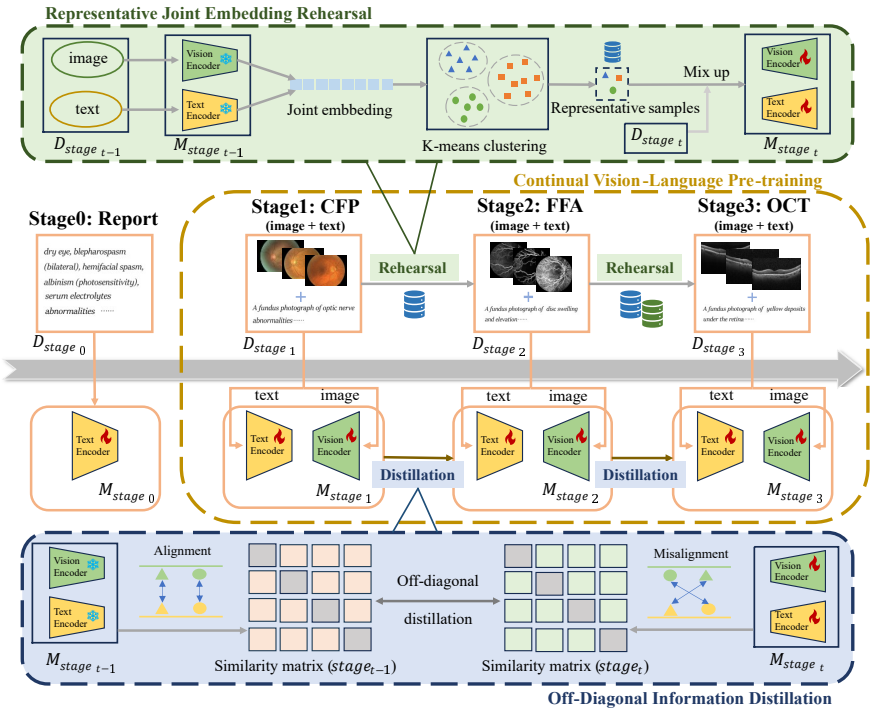

Fig. 1. An Overview of RetCoP. From left to right, the figure depicts the four-stage pre-training process, with stages $_1$, $_2$, and $_3$ performing continual vision-language pre-training on CFP, FFA, and OCT. The green section at the top shows the Representative Joint Embedding Rehearsal module, and the blue section at the bottom represents the Off-Diagonal Information Distillation module.

- We propose RetCoP, the first continual vision-language pre-training framework in the fundus domain, integrating image-text pairs from mainstream modalities into a unified foundation model.
- We introduce a rehearsal strategy based on representative image-text pairs and an off-diagonal information distillation strategy, effectively mitigating catastrophic forgetting in continual contrastive pre-training.
- Experimental results show that our method outperforms all the comparison approaches, achieving the best generalization performance and the lowest forgetting rate.

2 Method

2.1 Continual Vision-Language Pre-training Framework

As shown in Fig. 1, we propose RetCoP, a continual vision-language pre-training framework for fundus image analysis. The entire pre-training phase is divided

into four stages, marked from stage_0 to stage_3. In stage_0, we pre-train the text encoder with ophthalmology and general medical report text data by masked language modeling (MLM), providing a shared semantic base across varying image modalities for better text knowledge understanding and rich semantic representations for image-text alignment in subsequent contrastive pre-training. In the subsequent three-stage continual vision-language pre-training, we use image-text pairs from the CFP, FFA, and OCT modalities, which are introduced incrementally, in accordance with their decreasing data distribution in real clinical practice. Each stage independently learns a new modality representation.

During stage_1 to stage_3, we adopt a contrastive pre-training approach, following CLIP-like model [19] to align images and texts. The goal is to pull positive image-text pairs closer in the representation space while pushing negative pairs further apart. Suppose that there are N image-text pairs in each batch. After encoding and projection, the image feature representations and the text feature representations are denoted as: $I = [I_1, I_2, \ldots, I_N]$ and $T = [T_1, T_2, \ldots, T_N]$. The similarity matrix S is an $N \times N$ matrix, where each element S_{ij} represents the cosine similarity between the the feature vector I_i of i-th image and the feature vector T_j of the j-th text:

$$S_{ij} = \frac{I_i \cdot T_j}{\|I_i\| \|T_j\|}. \tag{1}$$

The contrastive loss $\mathcal{L}_{CLIP}$ is calculated based on the cosine similarity from both image-to-text and text-to-image. Specifically, it is formulated using the InfoNCE loss as follows:

$$\mathcal{L}_{CLIP} = -\frac{1}{2N} \sum_{i=1}^{N} \left[\log \frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N} \exp(S_{ij}/\tau)} + \log \frac{\exp(S_{ii}/\tau)}{\sum_{j=1}^{N} \exp(S_{ji}/\tau)} \right], \tag{2}$$

where S_{ii} is the similarity of the positive image-text pair, S_{ij} and S_{ji} are the similarities for negative pairs, τ is the temperature parameter.

In the continual pre-training scenario, an inevitable challenge is catastrophic forgetting, where knowledge acquired in previous stages is forgotten after training on a new stage. To mitigate forgetting, as shown in Fig. 1, we introduce two modules into our framework: representative joint embedding rehearsal and off-diagonal information distillation, constraining the model's continual updates by revisiting representative joint embedding and preserving the alignment of representations.

2.2 Representative Joint Embedding Rehearsal

A primary reason for model forgetting is losing access to data from previous stages during continual pre-training. To enable the model to revisit and review past knowledge, we construct a rehearsal mechanism based on representative joint embedding, preserving a portion of image-text pairs from previous modalities on earlier stages and mixing them with the current stage's modality data (see the green section in Fig. 1). First, to select representative samples while

considering both image and text features, we compute the joint embedding representation for each image-text pair. Specifically, for image embeddings I_i and text embeddings T_i, the joint embedding J_i is represented as:

$$J_i = S_{ii} \cdot I_i + (1 - S_{ii}) \cdot T_i, \tag{3}$$

where S_{ii} is the similarity score between I_i and T_i. If the image and text are highly aligned, the joint embedding relies more on the image information I_i; if weakly aligned, the text information T_i is enhanced in the joint embedding to ensure the completeness of the representation. Then, based on the computed joint embedding representations of the image-text pairs from the previous modality, we employ a k-means sampling strategy to select the most representative samples for constructing the rehearsal.

$$R = \bigcup_{k=1}^{K} \left\{ \arg\min_i \|J_i - c_k\| \right\}, \quad c_k = \frac{1}{|Q_k|} \sum_{i \in Q_k} J_i, \tag{4}$$

where c_k is the centroid of each cluster, Q_k represents the set of samples in the k-th cluster. For each cluster, a subset of samples closest to the centroid is selected and added to the rehearsal. K is the total number of clusters, and R is the final rehearsal constructed for this stage.

In the entire continual pre-training framework, we keep a slim and fixed-size rehearsal buffer, where data from all previous modalities are evenly distributed in the rehearsal buffer. The buffer is dynamically updated during pre-training as new modalities are added, preventing it from growing too large and imposing storage and computational burdens.

2.3 Off-Diagonal Information Distillation

In continual vision-language pre-training, a key factor affecting forgetting rate is the drift in the alignment between textual and visual features in the representation space. The textual and visual representations learned in previous stages may experience misalignment in new stages, leading to incorrect distance relationships between paired images and texts in the representation space.

Inspired by MOD-X [16], to further perceive and maintain the learned alignment between images and texts while revisiting previous knowledge, we introduce an off-diagonal information distillation method(see the blue section in Fig. 1). In the similarity matrix of each stage, the distribution of off-diagonal similarities reflects the current feature space. Thus, preserving off-diagonal information from past stages effectively maintains the alignment between image and text features across stages. S_{t-1} and S_t are the similarity matrices at stage_{t-1} and stage_t, respectively. Since the frozen model from the previous stage may misjudge image-text alignments in the current stage, we replace rows in the S_{t-1} (where diagonal similarity is not maximal) with corresponding values from the S_t to avoid misleading updates:

$$S_{t-1}^{(i,:)} = S_t^{(i,:)}; \quad \text{if} \quad \max(S_{t-1}^i) \neq i. \tag{5}$$

We use KL divergence to distill the similarity distribution from the previous stage's similarity matrix, thereby explicitly constraining the updates to the representations:

$$\mathcal{L}_{ODID}(S_t, S_{t-1}) = -\sum S_{t-1} \ln\left(\frac{S_t}{S_{t-1}}\right). \tag{6}$$

We incorporate the distillation loss into the contrastive loss $\mathcal{L}_{CLIP}$, resulting in the total loss function for RetCoP as follows, where λ is the weight for $\mathcal{L}_{ODID}$:

$$\mathcal{L}_{RetCoP} = \mathcal{L}_{CLIP} + \lambda \cdot \mathcal{L}_{ODID}. \tag{7}$$

3 Experiments

3.1 Experimental Setup

Pre-training Data. The text report in stage_0 is sourced from 34 public medical text datasets, spanning ophthalmology and general medicine [24]. From stage_1 to stage_3, we conduct continual vision-language pretraining using image-text pairs. For CFP modality, we use MM-Retinal[25] and flair dataset [21], which integrates 37 public fundus datasets covering 96 categories, totaling 193,678 pairs. For FFA modality, along with MM-Retinal[25], we utilize a subset of the FFA-IR [13], which covers 46 retinal diseases, resulting in a combined total of 701,098 pairs. For OCT modality, eleven public category-labeled OCT datasets and MM-Retinal [24] are leveraged, providing 183,817 pairs. Note that the categorical labels are mapped into textual prompts using prior knowledge.

Evaluation Data and Metrics. Disease classification was conducted on five multi-class fundus downstream datasets across CFP, FFA, and OCT modalities under zero-shot, linear probe, and CLIP-adapter [6] settings. These datasets cover many common retinal diseases like diabetic retinopathy, glaucoma, and age-related macular degeneration, among others. For CFP modality, the model was never trained on FIVES [10] but partially on ODIR [1] with three unseen categories withheld. Current-stage results demonstrate model plasticity, while previous-stage results reflect forgetting degree.

Methods for Comparison. RetCoP is compared with five popular continual learning methods: 1) SeqFT: The model is fine-tuned sequentially per stage. 2) LWF [14]: a regularization-based method using cross-entropy distillation. 3) EWC [11]: a regularization-based method leveraging Fisher matrix for weight constraints. 4) ER [4]: a replay-based method with reservoir sampling. 5) ICARL [20]: a replay-based method applying MoF sampling based on feature means.

Implementation Details. We employed ResNet-50 [8] initialized with ImageNet-1K as the vision encoder and BioClinicalBERT [2] as the text encoder. Images are resized to 512×512, and English text tokens are set to a length of 256. OCT images are resized with isotropic scaling and zero-padding. The evaluation uses five-fold cross-validation with averaged results. Training was conducted using the AdamW optimizer with a warm-up strategy on four NVIDIA A6000 GPUs, with a batch size of 24. The CLIP-like training follows [25].

Table 1. Performance Comparison on CFP Modality (FIVES and ODIR Datasets). Training FFA and OCT in stages $_2$ and $_3$ induces forgetting in CFP. The table shows downstream test results on CFP after all three stages, with forgetting rate relative to stage $_1$ in parentheses for stages $_2$ and $_3$.

FIVES Dataset							
Stage	Method	Zero-Shot		Linear Probe		ClipAdapter	
		ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)
1	-	58.9	88.6	82.6	96.1	81.6	95.0
2	RetCoP	**52.0(↓6.9)**	**82.2(↓6.4)**	**79.7(↓2.9)**	**94.5(↓1.6)**	**73.7(↓7.9)**	**91.7(↓3.3)**
	SeqFT	27.5(↓31.4)	60.7(↓27.9)	63.1(↓19.5)	85.2(↓10.9)	57.1(↓24.5)	82.7(↓12.3)
	LWF	44.2(↓14.7)	72.5(↓16.1)	72.4(↓10.2)	90.9(↓5.2)	64.6(↓17.0)	87.1(↓7.9)
	EWC	33.1(↓25.8)	65.6(↓23.0)	70.6(↓12.0)	89.5(↓6.6)	62.5(↓19.1)	84.2(↓10.8)
	ICARL	39.4(↓19.5)	65.0(↓23.6)	73.2(↓9.4)	90.9(↓5.2)	64.8(↓16.8)	86.5(↓8.5)
	ER	30.0(↓28.9)	74.5(↓14.1)	74.7(↓7.9)	92.7(↓3.4)	61.2(↓20.4)	85.9(↓9.1)
3	RetCoP	**45.9(↓13.0)**	**76.7(↓11.9)**	**77.6(↓5.0)**	**93.7(↓2.4)**	**72.3(↓9.3)**	**91.0(↓4.0)**
	SeqFT	21.9(↓37.0)	39.5(↓49.1)	57.4(↓25.2)	83.6(↓12.5)	55.6(↓26.0)	80.7(↓14.3)
	LWF	29.5(↓29.4)	59.3(↓29.3)	68.3(↓14.3)	88.1(↓8.0)	42.6(↓39.0)	70.7(↓24.3)
	EWC	21.5(↓37.4)	49.9(↓38.7)	64.0(↓18.6)	86.3(↓9.8)	48.7(↓32.9)	77.8(↓17.2)
	ICARL	29.2(↓29.7)	57.8(↓30.8)	66.4(↓16.2)	87.0(↓9.1)	39.7(↓41.9)	68.5(↓26.5)
	ER	26.1(↓32.8)	56.3(↓32.3)	73.5(↓9.1)	91.4(↓4.7)	62.1(↓19.5)	85.4(↓9.6)
ODIR Dataset							
Stage	Method	Zero-Shot		Linear Probe		ClipAdapter	
		ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)
1	-	80.0	93.8	89.0	97.9	89.0	96.6
2	RetCoP	**68.8(↓11.2)**	**90.0(↓3.8)**	**89.8(↑0.8)**	**97.9(0.0)**	**89.1(↑0.1)**	**97.7(↑1.1)**
	SeqFT	29.8(↓50.2)	48.2(↓45.6)	**89.8(↑0.8)**	96.4(↓1.5)	87.7(↓1.3)	96.0(↓0.6)
	LWF	41.5(↓38.5)	57.8(↓36.0)	89.0(0.0)	96.9(↓1.0)	87.5(↓1.5)	96.3(↓0.3)
	EWC	36.5(↓43.5)	70.7(↓23.1)	88.3(↓0.7)	96.0(↓1.9)	86.0(↓3.0)	95.3(↓1.3)
	ICARL	59.8(↓20.2)	85.8(↓8.0)	89.0(0.0)	97.1(↓0.8)	88.2(↓0.8)	96.5(↓0.1)
	ER	53.0(↓27.0)	73.4(↓20.4)	88.3(↓0.7)	96.6(↓1.3)	87.5(↓1.5)	96.8(↑0.2)
3	RetCoP	**56.2(↓23.8)**	**81.5(↓12.3)**	**91.3(↑2.3)**	**98.0(↑0.1)**	**90.7(↑1.7)**	**97.3(↑0.7)**
	SeqFT	30.0(↓50.0)	47.1(↓46.7)	73.5(↓15.5)	88.4(↓9.5)	69.3(↓19.7)	85.2(↓11.4)
	LWF	35.7(↓44.3)	51.9(↓41.9)	82.0(↓7.0)	92.0(↓5.9)	66.5(↓22.5)	81.9(↓14.7)
	EWC	32.7(↓47.3)	51.3(↓42.5)	77.3(↓11.7)	90.3(↓7.6)	75.8(↓13.2)	90.5(↓6.1)
	ICARL	34.3(↓45.7)	56.3(↓37.5)	79.8(↓9.2)	91.6(↓6.3)	66.8(↓22.2)	82.2(↓14.4)
	ER	38.8(↓41.2)	67.4(↓26.4)	88.7(↓0.3)	95.8(↓2.1)	87.7(↓1.3)	95.9(↓0.7)

Table 2. Performance Comparison on FFA Modality (MPOS Dataset). Stage $_3$ trains OCT, inducing FFA forgetting. The table shows downstream test results on FFA for stages $_2$ and $_3$, with values in parentheses indicating forgetting rate relative to stage $_2$.

Stage	Method	Zero-Shot		Linear Probe		ClipAdapter	
		ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)
2	RetCoP	**49.8**	**88.0**	**86.2**	**97.7**	80.7	**96.8**
	SeqFT	29.6	81.9	83.5	96.8	73.6	94.7
	LWF	44.7	80.4	85.5	97.2	**82.5**	96.5
	EWC	36.0	74.3	84.2	96.9	74.8	95.3
	ICARL	37.8	76.2	77.9	95.1	76.7	94.1
	ER	49.4	86.3	80.7	95.9	72.1	94.2
3	RetCoP	**25.4(↓24.4)**	**70.5(↓17.5)**	**83.0(↓3.2)**	**96.6(↓1.1)**	76.6(↓4.1)	95.0(↓1.8)
	SeqFT	18.0(↓11.6)	48.8(↓33.1)	76.9(↓6.6)	94.1(↓2.7)	54.0(↓19.6)	87.9(↓6.8)
	LWF	21.7(↓23.0)	69.8(↓11.1)	80.7(↓4.8)	95.5(↓1.7)	62.3(↓20.2)	92.5(↓2.2)
	EWC	17.9(↓18.1)	49.5(↓24.8)	70.8(↓13.4)	93.5(↓3.4)	50.7(↓24.1)	84.7(↓10.6)
	ICARL	20.8(↓17.0)	60.6(↓15.6)	77.7(↓0.2)	95.5(↓0.4)	67.2(↓5.5)	92.9(↓1.2)
	ER	23.2(↓26.2)	63.9(↓22.4)	81.1(↓0.4)	96.1(↓0.1)	**77.9(↑5.8)**	**95.3(↑1.1)**

Table 3. Performance Comparison on OCT Modality (ACC & AUC, %). The data in the table are averaged under zero-shot, linear probe and clip-adapter settings.

Dataset	Method					
	RetCoP	SeqFT	LWF	EWC	ICARL	ER
OCTID (ACC/AUC)	**71.6**/94.6	70.5/94.2	69.9/93.4	69.7/**95.5**	52.1/90.4	67.5/92.5
OCTDL (ACC/AUC)	**83.1**/**88.1**	82.4/87.4	81.7/83.8	82.6/87.7	71.3/86.8	80.0/84.0

Table 4. Ablation Study. The data in the table are averaged under zero-shot, linear probe and clip-adapter settings.

Rehearsal	ODID	Re-init	CFP_stage2		CFP_stage3		FFA_stage3	
			ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)
✓	✓	✓	71.3	89.3	64.1	86.4	58.7	85.7
	✓		62.8	81.2	51.6	73.7	56.5	83.5
✓			69.3	84.6	65.6	88.1	57.4	86.3
(ours) ✓	✓		**75.5**	**92.3**	**72.3**	**89.7**	**61.7**	**87.4**

3.2 Experimental Results and Analysis

Method Comparison. From Table 1, it is clear that RetCoP achieves the lowest forgetting rates and maintains superior performance across all stages and metrics in CFP modality, especially in zero-shot. For instance, on FIVES dataset [10], RetCoP shows only **6.9%** zero-shot ACC drop in stage_2 (vs. **31.4%** for SeqFT) and **13.0%** in stage_3 (vs. **37.0%** for SeqFT), demonstrating strong resistance to catastrophic forgetting. Notably, after training on FFA stage_2 and OCT stage_3, RetCoP retains **45.9%** zero-shot ACC for CFP in stage_3, while others (e.g., LWF: **29.5%**, EWC: **21.5%**) suffer severe degradation. On ODIR dataset [1], RetCoP not only mitigates forgetting (e.g., stage_3 zero-shot AUC: ↓**12.3%** vs. SeqFT's ↓**46.7%**) but also even improves some metrics (e.g., ClipAdapter AUC: ↑**1.1%** in stage_2).

As shown in Table 2, on the MPOS dataset [23] (FFA modality), RetCoP achieves the least forgetting in stage_3, with only **3.2%** ACC decline in linear probe and **1.1%** AUC drop – significantly lower than others like SeqFT (↓**6.6%** ACC/↓2.7% AUC) and EWC (↓**13.4%** ACC/↓3.4% AUC). While ER shows accidental improvement in ClipAdapter (↑5.8% ACC), its severe forgetting in zero-shot (↓**26.2%** ACC) reveals instability. RetCoP achieves these results by effectively balancing cross-modal knowledge through rehearsal and distillation, ensuring stable performance on prior tasks while adapting to new modalities. The results in FFA (Table 2) and OCT (Table 3, on OCTID [7] and OCTDL [12] datasets) experiments demonstrate that RetCoP not only effectively mitigates forgetting but also achieves the best performance in learning FFA and OCT modalities during stage_2 and stage_3, showcasing strong plasticity.

Ablation Studies. To validate the effectiveness of each module in mitigating catastrophic forgetting, we designed ablation experiments using the control variable method (Table 4). The results show that removing the rehearsal or distillation component increases the model's average forgetting rate in CFP and FFA modalities, highlighting their critical role in facilitating knowledge retention and

mitigating catastrophic forgetting. Additionally, we found that reinitializing the text encoder with stage_0 weights during incremental pre-training led to higher forgetting rates than using the previous stage's weights.

4 Conclusion

In this paper, we propose RetCoP, the first continual pretraining framework in fundus domain, learning incremental imaging modalities. Experiments demonstrate that, by designing the rehearsal strategy and off-diagonal information distillation, RetCoP effectively achieves optimal performance and mitigates catastrophic forgetting during continual pre-training. We hope this work will inspire further research for incremental pretraining in dynamic clinical environments.

Acknowledgments. This work was partially supported by the National Natural Science Foundation of China (Nos. 62476054, and 62172228), and the Fundamental Research Funds for the Central Universities of China.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. International competition on ocular disease intelligent recognition (2019), `https://odir2019.grand-challenge.org`
2. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
3. Badar, M., Haris, M., Fatima, A.: Application of deep learning for retinal image analysis: A review. Computer Science Review **35**, 100203 (2020)
4. Chaudhry, A., Rohrbach, M., Elhoseiny, M., Ajanthan, T., Dokania, P., Torr, P., Ranzato, M.: Continual learning with tiny episodic memories. In: Workshop on Multi-Task and Lifelong Reinforcement Learning (2019)
5. Du, J., Guo, J., Zhang, W., Yang, S., Liu, H., Li, H., Wang, N.: Ret-clip: A retinal image foundation model pre-trained with clinical diagnostic reports. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 709–719. Springer (2024)
6. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)
7. Gholami, P., Roy, P., Parthasarathy, M.K., Lakshminarayanan, V.: Octid: Optical coherence tomography image database. Computers & Electrical Engineering **81**, 106532 (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. He, X., Zhou, Y., Wang, B., Cui, S., Shao, L.: Dme-net: Diabetic macular edema grading by auxiliary task learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 788–796. Springer (2019)

10. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Scientific data **9**(1), 475 (2022)
11. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
12. Kulyabin, M., Zhdanov, A., Nikiforova, A., Stepichev, A., Kuznetsova, A., Ronkin, M., Borisov, V., Bogachev, A., Korotkich, S., Constable, P.A., et al.: Octdl: Optical coherence tomography dataset for image-based deep learning methods. Scientific data **11**(1), 365 (2024)
13. Li, M., Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., Chen, X., Liu, Z., Pan, C., Li, M., et al.: Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In: Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2) (2021)
14. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
15. Li, Z., Song, D., Yang, Z., Wang, D., Li, F., Zhang, X., Kinahan, P.E., Qiao, Y.: Visionunite: A vision-language foundation model for ophthalmology enhanced with clinical knowledge. arXiv preprint arXiv:2408.02865 (2024)
16. Ni, Z., Wei, L., Tang, S., Zhuang, Y., Tian, Q.: Continual vision-language representation learning with off-diagonal information. In: International Conference on Machine Learning. pp. 26129–26149. PMLR (2023)
17. Pan, L., Chen, X.: Retinal oct image registration: methods and applications. IEEE reviews in biomedical engineering **16**, 307–318 (2021)
18. Qiu, J., Wu, J., Wei, H., Shi, P., Zhang, M., Sun, Y., Li, L., Liu, H., Liu, H., Hou, S., et al.: Visionfm: a multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence. arXiv preprint arXiv:2310.04992 (2023)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
20. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
21. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J., Ayed, I.B.: A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. Medical Image Analysis **99**, 103357 (2025)
22. Tasai, R., Li, G., Togo, R., Tang, M., Yoshimura, T., Sugimori, H., Hirata, K., Ogawa, T., Kudo, K., Haseyama, M.: Continual self-supervised learning considering medical domain knowledge in chest ct images. arXiv preprint arXiv:2501.04217 (2025)
23. Wang, H., Xing, Z., Wu, W., Yang, Y., Tang, Q., Zhang, M., Xu, Y., Zhu, L.: Non-invasive to invasive: Enhancing ffa synthesis from cfp with a benchmark dataset and a novel network. In: Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine. pp. 7–15 (2024)
24. Wu, R., Su, N., Zhang, C., Ma, T., Zhou, T., Cui, Z., Tang, N., Mao, T., Zhou, Y., Fan, W., et al.: Mm-retinal v2: Transfer an elite knowledge spark into fundus vision-language pretraining. arXiv preprint arXiv:2501.15798 (2025)

25. Wu, R., Zhang, C., Zhang, J., Zhou, Y., Zhou, T., Fu, H.: Mm-retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 722–732. Springer (2024)
26. Yan, S., Hong, L., Xu, H., Han, J., Tuytelaars, T., Li, Z., He, X.: Generative negative text replay for continual vision-language pretraining. In: European Conference on Computer Vision. pp. 22–38. Springer (2022)
27. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Wu, Q., Xia, Y.: Continual self-supervised learning: Towards universal multi-modal medical data representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11114–11124 (2024)
28. Zheng, Z., Ma, M., Wang, K., Qin, Z., Yue, X., You, Y.: Preventing zero-shot transfer degradation in continual learning of vision-language models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 19125–19136 (2023)
29. Zhu, H., Wei, Y., Liang, X., Zhang, C., Zhao, Y.: Ctp: Towards vision-language continual pretraining via compatible momentum contrast and topology preservation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22257–22267 (2023)