Continually Tuning a Large Language Model for Multi-domain Radiology Report Generation

Yihua Sun[1], Hee Guan Khor[1], Yuanzheng Wang[2], Zhuhao Wang[1], Hongliang Zhao[2], Yu Zhang[2], Longfei Ma[1], Zhuozhao Zheng[2], and Hongen Liao[1(✉)]

[1] School of Biomedical Engineering, Tsinghua University, Beijing, China
liao@tsinghua.edu.cn

[2] Department of Radiology, Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua University, Beijing, China

Abstract. Large language models (LLMs) have demonstrated potential across various tasks, including vision-language applications like chest X-ray (XR) report generation (RG) in healthcare. Recent RG approaches focus on optimizing model performance for a single dataset with a single XR modality, often neglecting the critical area of computed tomography (CT) report generation. The challenge is compounded by medical datasets being isolated across different centers, making comprehensive collection difficult. Furthermore, LLMs trained on datasets sequentially can experience catastrophic forgetting. In this paper, we move beyond conventional approaches of training on a single dataset, and focus on improving the overall performance on sequentially collected multi-center datasets. We incorporate four datasets with diverse languages and image modalities for the experiments. Our approach utilizes a minimal number of task-specific learnable weights within an LLM-based RG method for each domain, maintaining the majority of weights frozen to avoid forgetting. Utilizing LLMs' multilingual generalizability, we align models and facilitate knowledge sharing through a multi-label supervised contrastive loss within the LLM hidden space. We design a 2D-3D adapter for the image encoder to transfer from XR to CT RG tasks. A CT disease graph is established for transferring knowledge from XR to CT RG tasks, using CT's most relevant XR disease class centers in a triplet loss. Extensive experiments validate our design.

Keywords: Continual learning · Large language model · Multi-domain · Multi-modality · Parameter efficient fine-tuning · Report generation.

1 Intorduction

Integrating various modalities and tasks into a unified system offers a promising avenue toward achieving medical artificial general intelligence. Large language models (LLMs) trained on extensive textual datasets have shown impressive results [3,5,8,31,32,38]. Recent advancements have extended LLM applications to vision-language tasks [4,16], including chest X-ray (XR) report generation (RG), enhancing clinical efficiency and reducing radiologists' workload [23,30,36,37].

However, current practices in the medical field primarily involve fine-tuning LLMs for specific applications [23,30,36,37]. Medical data is isolated in different centers, presenting challenges in accessing them at large scale all at once. This isolation prevents access to comprehensive datasets, challenging the tuning of a unified LLM for RG. On the other hand, the sequential gathering of multi-center data poses a risk of catastrophic forgetting for LLM once it learns a new task [20]. Moreover, current RG methods focus largely on XR [22], with a noticeable lack of focus on the clinically significant 3D computed tomography (CT).

The advent of LLMs presents an opportunity to create a versatile model for RG, tailored to the diverse needs of medical centers. Moving beyond conventional approaches of training on a single dataset, there is a critical need for RG models that can continually learn and be optimized for all domains without forgetting. Yet, most existing continual learning strategies concentrate on classification tasks [35], leaving the more challenging RG task unaddressed. Our research aims to fill this gap by developing continual learning methods for multi-domain radiology RG, targeting both technical innovation and clinical impact.

We propose a novel paradigm, namely **C**ontinually tuning for **M**ulti-domain **R**adiology **R**eport **G**eneration based on **LLM** (**CMRG-LLM**), aimed at optimizing the performance across sequentially acquired multi-center RG datasets and transferring knowledge from XR to CT RG. The major challenges are: 1) Addressing domain shift concerning both images and the linguistic style of reports; 2) Facilitating efficient knowledge transfer across tasks, notably from XR to CT RG; 3) Meeting clinical requirements that utilize varying numbers of input images; 4) Preventing the forgetting of previous tasks.

We leverage the strong generalizability of LLMs, with a focus on multilingual capabilities [10] to produce consistent embedding across languages. A limited number of weights is learned for each domain to bridge disparities in visual and linguistic contexts. Subsequently, we prompt a frozen LLM to produce desired outputs for various domains and prevent forgetting. We utilize a multi-labeled supervised contrastive loss to cluster features, fostering knowledge transfer across tasks by leveraging previous disease class centroids within the latent space of the LLM. Furthermore, alongside the existing XR disease graph [11], we create a graph of common diseases in CT and link them to their most relevant XR diseases. For transitioning from XR to CT, we utilize a lightweight 2D-3D adapter to manage dimensional expansion and transfer knowledge using the class center of CT's most relevant disease on XR. Extensive experiments validate our design.

2 Methodology

2.1 Problem Formulation

A sequence of datasets $\{D_t\}_{t=1}^{T}$, $D_t = \{(X_i^t, y_i^t)\}_{i=1}^{N_t}$, is collected from multi-center clinical sources, encompassing diverse image modalities. In each domain t, $X_i^t = \{x_1^t, \cdots, x_{k_i}^t\}$ represents a set of images, and y_i^t is the corresponding medical report. Given that a radiologist may capture one or multiple image(s)

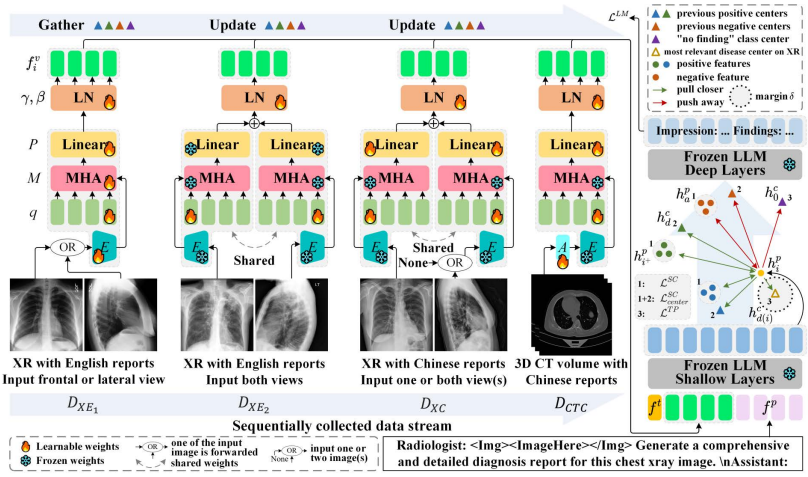

Fig. 1. We focus on sequentially collected multi-center, multi-modality datasets. (bottom of the figure). Once trained on the initial dataset, CMRG-LLM acquires a minimal set of parameters to adapt to new datasets and prevent forgetting. Disease features are clustered in the LLM's hidden space, facilitating knowledge transfer by aligning features with the class center from previous datasets (right-hand side of the figure).

(e.g., both frontal and lateral chest XR) for a single report, the image count k_i in X_i^t can vary. Each X_i^t is associated with a multi-hot disease label e_i.

For multi-domain RG experiments, we collect four datasets featuring diverse report styles and image modalities as follows: 1) D_{XE_1}: A large-scale public dataset containing chest **X**R images paired with **E**nglish reports. 2) D_{XE_2}: A smaller dataset contains chest **X**R images and **E**nglish reports, which specifically include paired frontal and lateral views. 3) D_{XC}: A clinical dataset consists of chest **X**R images and **C**hinese reports collected from real clinical settings, where images may be singular or paired to generate a report. 4) D_{CTC}: A clinical dataset comprises 3D chest **CT** volumes and corresponding reports in **C**hinese. Further details regarding these datasets are available in Section 3.1.

2.2 Prompt Construction

We start with a large public dataset and continually train a model to address various clinical needs without forgetting. We conduct experiments with the order D_{XE_1}→D_{XE_2}→D_{XC}→D_{CTC}, reflecting the practice of model development from online to clinical datasets and from simpler XR tasks to more intricate CT tasks.

Large Language Model. It is observed that multilingual LLMs can produce consistent latent embedding across languages [10]. In light of this, we utilize a multilingual LLM as the backbone (Fig. 1). The parameters Ω in the LLM are frozen, to ensure efficient tuning while maintaining pre-trained generalizability. An instruction prompts the LLM to activate its knowledge for RG, depicted in Fig. 1 and supplementary material. A tokenizer embeds the instruction into f^p.

Learnable Prompts. A proficient continual RG learner can adeptly manage the domain gap in report stylistics/language and image representation. The overall framework is depicted in Fig. 1. We design a domain token f^t, sized $1 \times C$, shared across all images within a domain to capture the style of reports, where C is the channel dimension. Besides, a learnable query q is utilized to address image discrepancies across domains by interacting with the features of X_i^t.

To obtain an image's embedding sequence f_k of length L_f, we employ a 2D visual encoder $E(\cdot;\theta)$, where $f_k = E(x_k^t;\theta)$ and θ are the learnable parameters. Inspired by [16] for bridging the visual-lingual gap, we adopt q (shaped $L_q \times C$) as query and f_k as key/value in a multi-head attention (MHA) [33] block $M(\cdot;\mu)$ to aggregate visual information. The aggregated image embedding is: $f_k^q = M(Q,K,V;\mu) = M(q,f_k,f_k;\mu)$, where f_k^q matches the shape of q with $L_q < L_f$. This mechanism enables M to distill the most informative visual features from f_k for transformation into textual representations. The feature f_k^q is projected by a linear layer $P(\cdot;\sigma)$ to match the LLM's hidden size.

Even multiple input images are present, we derive a single feature f_i by averaging their projected features: $f_i = \frac{1}{k_i}\sum_{k=1}^{k_i} P(f_k^q;\sigma)$, using shared parameters σ. This feature f_i undergoes refinement through layer normalization (LN) [2], employing parameters γ and β, to generate the visual prompt f_i^v.

The final prompt for LLM is the concatenation of f^t, f_i^v, and f^p. The training goal is to maximize the output's *log* probability in an auto-regressive manner [26]:

$$\mathcal{L}^{LM}(f^t,\theta,\mu,q,\sigma,\gamma,\beta,\alpha) = -\sum_{(X_i^t,y_i^t)\in D_t} log\ p(y_i^t|[f^t,f_i^v,f^p],\Omega^*,\theta,\mu,q,\sigma,\gamma,\beta,\alpha), \quad (1)$$

where "$*$" indicates that the weights are frozen and α is optional for 2D-3D adaptation that will be introduced in Section 2.3. For the the initial domain D_{XE_1}, we tune all parameters with loss $\mathcal{L}_{XE_1}^{LM} = \mathcal{L}^{LM}(f^t,\theta,\mu,q,\sigma,\gamma,\beta)$.

2.3 Tuning Strategy

For the initial training on D_{XE_1}, following [37], we let $|X_i^t| = 1$ to input either frontal or lateral view of chest XR, ensuring E could effectively process both views. On D_{XE_2}, $|X_i^t| = 2$, indicating inputs comprising both views. On D_{XC}, $|X_i^t| = 1$ or 2, reflecting real clinical applications. On D_{CTC}, $|X_i^t| = 1$.

Tunable Parameters. After being trained on D_{XE_1}, E gains the capability to derive valuable insights from XR images, facilitated by M linking these images to LLM. To prevent knowledge loss, we freeze E and M and introduce unique parameters for f^t and q across different domains. This strategy helps in reducing the disparities in report styles and image inputs observed across domains.

We tailor the optimal prompt for each domain through distinct parameters for P and γ/β in LN. The narrower domain gap between the transition from D_{XE_1} to D_{XE_2}, where both datasets are in English, allows us to maintain a fixed P. The language modeling losses on D_{XE_2} and D_{XC} are: $\mathcal{L}_{XE_2}^{LM}(f^t,q,\gamma,\beta) = \mathcal{L}^{LM}(f^t,\theta^*,\mu^*,q,\sigma^*,\gamma,\beta)$, $\mathcal{L}_{XC}^{LM}(f^t,q,\sigma,\gamma,\beta) = \mathcal{L}^{LM}(f^t,\theta^*,\mu^*,q,\sigma,\gamma,\beta)$.

Forward Knowledge Alignment through LLM. We propose aligning models with disease labels during continual learning within the latent space of LLM, facilitating forward knowledge transfer through class centers. Given an input set $[f^t, f_i^v, f^p]$, we denote its hidden features in the LLM as h_i, with a length of L_h. We calculate feature h_i^p using a projection layer P_h: $h_i^p = P_h\left(\frac{1}{L_h}\sum_{l=1}^{L_h} h_i(l,:)\right)$.

Initially, lacking prior information, we group h_i^p into N_{XR} disease classes using supervised contrastive loss [14]. Considering the possibility of multiple disease labels per image, we adjust for a multi-label context as follows:

$$\mathcal{L}^{SC} = \sum_i \frac{1}{|S(h_i^p)|} \sum_{h_{i^+}^p \in S(h_i^p)} log \frac{exp\left(\langle h_i^p, h_{i^+}^p \rangle / \tau\right)}{\sum_{a \in B \setminus \{i\}} exp\left(\langle h_i^p, h_a^p \rangle / \tau\right)}, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is inner product, $\tau \in \mathbb{R}^+$ is a scalar temperature parameter, B is a set of batch indices. $S(h_i^p)$ includes positive samples sharing at least one label with h_i^p. In this case, samples with multiple labels will be pulled to all the positive groups, and helps to pull these groups closer. Given the higher intrinsic correlations among co-existing diseases, $\mathcal{L}^{SC}$ model these correlations by pulling the classes closer using the multi-labeled samples (Fig. 1).

For $D_{XE_1} \rightarrow D_{XE_2} \rightarrow D_{XC}$, we calculate and update feature centers h_d^c for each disease class, using only features associated with a single label. During training on D_{XE_2}, we group features by their labels and align them with corresponding class centers from the previous dataset to promote knowledge transfer. We define $\mathcal{L}_{center}^{SC}$ similar to Eq. (2), but update $S(h_i^p)$ to $S'(h_i^p) = S(h_i^p) \cup \{h_d^c | e_i(d) = 1\}$, where $e_i(d) = 1$ indicates that the sample X_i^t have a label of d-th disease.

From XR to CT. Given a 3D CT with shape $H \times W \times Z$, it offers broader disease detection capabilities compared to XR. Instead of developing a 3D encoder for CT from scratch and discarding XR-derived insights, we propose a lightweight 2D-3D adapter, $A(\cdot\,; \alpha)$, with only two convolution layers. We reposition the longitudinal dimension Z to the channel dimension, compressing it into 3 channels using 2D convolutions. It processes CT's in-plane data via 2D convolution, merging it across planes to encapsulate 3D information as it condenses the channels. The resulting $3 \times H \times W$ features are fed into the pre-trained frozen E (Fig. 1). A is trained using $\mathcal{L}_{CTC}^{LM}(f^t, q, \sigma, \gamma, \beta, \alpha) = \mathcal{L}^{LM}(f^t, \theta^*, \mu^*, q, \sigma, \gamma, \beta, \alpha)$.

We align the common disease between CT and XR with the previous feature centers on XR using $\mathcal{L}_{center}^{SC}$. For diseases unique to CT, features are drawn towards the most relevant disease class center, $h_{d(i)}^c$, from XR (supplementary material) and repelled from the "no findings" center, h_0^c, using a triplet loss [29]: $\mathcal{L}^{TP} = max(\langle h_{d(i)}^c, h_i^p \rangle - \langle h_0^c, h_i^p \rangle + \delta, 0)$, where δ is the margin parameter. This enhances the transfer of disease correlation knowledge from XR to CT. For diseases unique to CT that should not be strictly pulled to XR features, δ allows the relaxation for the features to explore around, preventing strict alignment.

Losses. The final losses are as follows: 1) On D_{XE_1}, $\mathcal{L} = \mathcal{L}_{XE_1}^{LM} + \lambda_1^c \mathcal{L}^{SC}$; 2) On D_{XE_2}, $\mathcal{L} = \mathcal{L}_{XE_2}^{LM} + \lambda_1^c \mathcal{L}_{center}^{SC}$; 3) On D_{XC}, $\mathcal{L} = \mathcal{L}_{XC}^{LM} + \lambda_1^c \mathcal{L}_{center}^{SC}$; 4) On D_{CTC}, $\mathcal{L} = \mathcal{L}_{CTC}^{LM} + \lambda_1^c \mathcal{L}_{center}^{SC} + \lambda_2^c \mathcal{L}^{TP}$, where λ_1^c, λ_2^c are hyper-parameters.

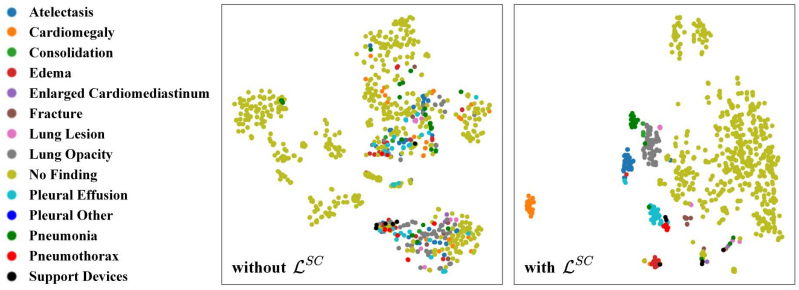

Fig. 2. The t-SNE [21] visualizations of LLM hidden features. LLM shows potential by distinguishing "no finding" and others with only $\mathcal{L}^{LM}$. We further shape the feature space with $\mathcal{L}^{SC}$ and disease class centers are utilized for knowledge transferring.

3 Experiments

3.1 Experimental Setup

Datasets. Four datasets are included in the study: 1) MIMIC-CXR (D_{XE_1}): A public dataset comprises chest XR images and English reports [9,12,13]. We collect 238k images and follow the official split. 2) IU-Xray (D_{XE_2}): A public dataset includes paired frontal and lateral chest XR images and English reports [7]. We collect 5.9k images and follow the 7:1:2 train/validation/test set split in [17]. 3) D_{XC}: A clinical dataset collected from Beijing Tsinghua Changgung Hospital (BTCH), containing 5.7k chest XR images and Chinese reports, encompassing both paired and singular frontal and lateral views. 4) D_{CTC}: A clinical dataset collected from BTCH, comprising 4.3k chest CT scans and Chinese reports.

We employ a random 7:1:2 split to partition D_{XC} and D_{CTC} into training, validation, and test sets. For all datasets, we filter data containing both impression and findings sections, and predict both sections. The reports are labeled based on CheXpert labeler [11,25].

Implementation Details and Evaluation Metrics. The XR images are downsampled to 224×224, and the CT volumes are downsampled to 224×224×64 with a spacing of $1.5 \times 1.5 \times 5$ mm^3. We utilize Qwen [3] with 7B parameters as the LLM backbone. Encoder E is a Swin Transformer [19] pre-trained on ImageNet [28]. The Chinese reports are segmented by jieba [1]. More details are in the supplementary material. We use BLEU-n [24], ROUGE [18] and CIDEr [34] for evaluation. We quantify the overall performance using the average score across current and previous tasks. We report score $v_t = \frac{1}{t}\sum_{s=1}^{t} v_{s,t}$, where $v_{s,t}$ is the metric v of a model trained on t-th task evaluated on the test set of s-th task.

3.2 Quantitative and Qualitative Evaluations

Visualizations of LLM Hidden Space. The t-SNE visualizations (Fig. 2) demonstrate that without the contrastive loss $\mathcal{L}^{SC}$, there is overlap among clusters corresponding to diseased classes. However, clusters associated with the

Table 1. Comparison with SOTA continual learning methods. CMRG-LLM outperforms other methods. The scores of R2genGPT are quoted from their original paper.

Methods	BLEU-3			BLEU-4			ROUGE			CIDEr		
t-th task	$t=2$	$t=3$	$t=4$	$t=2$	$t=3$	$t=4$	$t=2$	$t=3$	$t=4$	$t=2$	$t=3$	$t=4$
R2GenGPT [37]	0.207	-	-	0.154	-	-	0.337	-	-	0.354	-	-
Per-task FT	0.209	0.355	0.353	0.152	0.308	0.307	0.358	0.495	0.480	0.343	1.533	1.334
SeqFT	0.167	0.220	0.127	0.117	0.209	0.107	0.324	0.279	0.190	0.295	1.307	0.186
Replay	0.220	0.355	0.364	0.164	0.309	0.318	0.362	0.492	0.485	**0.440**	**1.543**	**1.360**
EWC [15]	0.173	0.265	0.125	0.123	0.240	0.107	0.331	0.365	0.190	0.314	1.346	0.216
DER [6]	0.145	0.218	0.282	0.105	0.209	0.249	0.281	0.264	0.376	0.338	1.301	1.155
ProgPrompt [27]	0.210	0.359	0.358	0.153	0.312	0.313	0.354	0.491	0.475	0.353	1.531	1.305
CMRG-LLM	**0.229**	**0.377**	**0.376**	**0.169**	**0.328**	**0.330**	**0.374**	**0.507**	**0.489**	0.290	1.529	1.348

"no finding" class show some degree of separation due to the auto-regressive report generation loss $\mathcal{L}^{LM}$. This observation suggests LLM's capability to encode disease-related knowledge. Incorporating the $\mathcal{L}^{SC}$ loss, which pulls samples with shared labels closer and pushes others apart, enhances the representation.

Diseases that frequently co-occur exhibit a stronger intrinsic correlation. The $\mathcal{L}^{SC}$ captures this correlation by pulling the feature clusters of frequent co-occurring classes closer, using the multi-labeled samples, while pushing them apart otherwise. As shown in Fig. 2, "cardiomegaly" is distinct from lung and pleural diseases. Subsequently, we encourage forward knowledge transfer by aligning the newly learned feature space with the previous class centers.

Quantitative Results. We compare our method to conventional methods: 1) Per-task FT: Fine-tuning and caching an independent parameter for each task. 2) SeqFT: Sequentially fine-tuning all parameters across a sequence of tasks. 3) Replay: Fine-tuning with a memory buffer and replay sample from old tasks. We also compare state-of-the-art (SOTA) general continual learning methods EWC [15], DER [6] and progressive prompt (ProgPrompt) [27]. ProgPrompt sequentially concatenating new learnable prompts for each task to LLM. Note that we consider all the parameters for tuning except the LLM is kept frozen.

As shown in Table 1, after being trained on the 4-th task (D_{CTC}), the performance of SeqFT dropped much lower. This result indicates that LLM for RG tasks suffers from catastrophic forgetting, which hinders its expansion in real clinical applications and emphasizes the necessity of our research. In Table 1, our methods outperform SOTA methods, indicating effectiveness in learning and transferring knowledge on sequential tasks. Per-task FT optimizes its performance for a single dataset, discarding the knowledge in other datasets, thus limiting its performance.

Since Per-task FT generates sub-optimal results, it is not true that the more parameters tuned the better the performance can be. Ideally, parameters with domain-specific knowledge should be updated, while those with domain-invariant knowledge should remain unchanged to facilitate knowledge transfer and mitigate forgetting. We then evaluate if each parameter tuned is effective for overall performance by freezing one of them. As shown in Table 2, each tuned parameter contributes to the final score. P is crucial for generating prompts for each domain and triggering the desired outputs. The empirical results suggest that q

Table 2. Experimental results of how each learnable parameters contribute to the final score. Also, the effect of $\mathcal{L}^{SC}$ and previous class center h_d^c is evaluated.

q	f^t	P	$\mathcal{L}^{SC}$	h_d^c	BLEU-3			BLEU-4			ROUGE			CIDEr		
t-th task					$t=2$	$t=3$	$t=4$	$t=2$	$t=3$	$t=4$	$t=2$	$t=3$	$t=4$	$t=2$	$t=3$	$t=4$
✓	✓		✓		0.228	0.364	0.358	0.168	0.315	0.311	**0.374**	0.495	0.472	**0.302**	1.442	1.191
✓		✓	✓		0.227	0.370	0.370	0.167	0.321	0.324	0.373	0.505	0.488	0.290	1.497	1.318
	✓	✓	✓		**0.229**	0.374	0.373	0.168	0.325	0.326	0.373	0.506	**0.490**	0.296	1.501	1.305
✓	✓	✓			0.216	0.361	0.359	0.155	0.311	0.313	0.361	0.495	0.477	0.263	1.455	1.268
✓	✓	✓	✓		0.228	0.375	0.373	0.168	0.326	0.326	**0.374**	**0.507**	**0.490**	**0.302**	1.521	1.317
✓	✓	✓	✓	✓	**0.229**	**0.377**	**0.376**	**0.169**	**0.328**	**0.330**	**0.374**	**0.507**	0.489	0.290	**1.529**	**1.348**

Fig. 3. The outputs and corresponding translations (key information is in color). SeqFT suffers from forgetting and cannot generalize well to CT RG (highlighted in yellow). The precise location/size for CT RG is still difficult to generate (highlighted in green).

and f^t capture domain-specific discrepancies in image and report style. Table 2 further demonstrates that the multi-label contrastive loss $\mathcal{L}_{SC}$ and the previous class center h_d^c enhance knowledge transfer, resulting in improved outcomes.

Case Study. After the models have sequentially learned up to the 4-th task D_{CTC}, we evaluate their performance on D_{XE_1} and D_{CTC}, as shown in Fig. 3. SeqFT incorrectly identifies the language style and erroneously generates CT findings from an XR image, as highlighted in yellow. In contrast, our method produces a more accurate CT report, showcasing enhanced capability in transferring knowledge from XR to CT and accurately detecting CT-specific findings (e.g., parenchymal band, nodule).

4 Conclusion and Discussion

In this paper, we propose a novel paradigm for continual learning in RG using LLM, moving beyond previous strategies that target a single dataset. We employ

minimal task-specific learnable parameters to adjust to new domains, addressing variations in image and report styles. We enhance knowledge transfer across domains by incorporating disease class centers. Additionally, we present a CT disease graph linked to the most relevant XR disease, facilitating effective cross-modality transfer with a 2D-3D adapter. The limitation of this work is that long reports and precise location/size of diseases are still very challenging to generate (Fig. 3). This may be improved by incorporating human-LLM interactions with manual textual or visual prompts. We hope our work can bring new insights to the community in the era of large/foundation models.

Acknowledgments. The authors acknowledge supports from National Key Research and Development Program of China (2022YFC2405200), National Natural Science Foundation of China (82027807, U22A2051), Institute for Intelligent Healthcare, Tsinghua University (2022ZLB001), and Tsinghua-Foshan Innovation Special Fund (2021THFS0104). The authors would like to thank Fang Chen for valuable discussions.

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

References

1. jieba. https://github.com/fxsjy/jieba
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Bai, J., Bai, S., et al.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901 (2020)
6. Buzzega, P., Boschini, M., Porrello, A., Abati, D., CALDERARA, S.: Dark experience for general continual learning: a strong, simple baseline. In: Advances in Neural Information Processing Systems. vol. 33, pp. 15920–15930 (2020)
7. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., et al.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (07 2015)
8. Du, Z., Qian, Y., et al.: GLM: General language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 320–335 (2022)
9. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., et al.: PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation **101**(23), E215–20 (Jun 2000)
10. Hu, J., Yao, Y., Wang, C., Wang, S., Pan, Y., Chen, Q., Yu, T., Wu, H., et al.: Large multilingual models pivot zero-shot multimodal learning across languages. In: The Twelfth International Conference on Learning Representations (2024)
11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 590–597 (Jul 2019)

12. Johnson, A., Pollard, T., Mark, R., Berkowitz, S., Horng, S.: MIMIC-CXR Database (version 2.0.0). PhysioNet (2019), https://doi.org/10.13026/C2JT1Q.
13. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**(1), 317 (2019)
14. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 18661–18673 (2020)
15. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
16. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
17. Li, Y., et al.: Hybrid retrieval-generation reinforced agent for medical image report generation. In: Advances in Neural Information Processing Systems. vol. 31 (2018)
18. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (2021)
20. Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., Zhang, Y.: An empirical study of catastrophic forgetting in large language models during continual fine-tuning. arXiv preprint arXiv:2308.08747 (2023)
21. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
22. Monshi, M.M.A., Poon, J., Chung, V.: Deep learning in generating radiology reports: A survey. Artificial Intelligence in Medicine **106**, 101878 (2020)
23. OpenMEDLab: Xraypulse. https://github.com/openmedlab/XrayPULSE (2023)
24. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. p. 311–318. ACL '02 (2002)
25. Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., Lu, Z.: NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Summits Transl. Sci. Proc. **2017**, 188–196 (May 2018)
26. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)
27. Razdaibiedina, A., Mao, Y., Hou, R., Khabsa, M., Lewis, M., Almahairi, A.: Progressive prompts: Continual learning for language models. In: The Eleventh International Conference on Learning Representations (2023)
28. 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**(3), 211–252 (2015)
29. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
30. Thawkar, O., Shaker, A., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.S.: Xraygpt: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971 (2023)

31. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
32. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
34. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015)
35. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: Theory, method and application. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
36. Wang, R., Duan, Y., Li, J., Pang, P., Tan, T.: Xrayglm: The first chinese medical multimodal model that chest radiographs summarization. https://github.com/WangRongsheng/XrayGLM (2023)
37. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology **1**(3), 100033 (2023)
38. Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al.: GLM-130b: An open bilingual pre-trained model. In: The Eleventh International Conference on Learning Representations (2023)