Diagnose Like a Doctor: Misdiagnosis-Aware Structured Memory for Continual Learning of Medical Vision-Language Models

Chendong Qin[1], Shuai Huang[2], and Yifei Yao[3]

[1] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] College of Computer Science and Artificial Intelligence, Fudan University, Shanghai, China
[3] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

Abstract. Multimodal biomedical Vision-Language Models (VLMs) confront a core dilemma in Continual Learning (CL): fine-grained feature confusion and blurred decision boundaries arising from the distinct data distribution characteristics of medical imagery. To address this challenge, we propose a novel Misdiagnosis-Aware Structured Memory framework that explicitly embeds misdiagnosis lessons into the retrieval topology, enabling the model to perform diagnostics akin to an experienced clinician. Specifically, we build a hierarchical Pitfall Graph that connects medical entities to their definitions, common mimics, and exclusion criteria, and integrate it into a Retrieval-Augmented Generation (RAG) system to guide fine-tuning via dynamic knowledge retrieval. To mitigate catastrophic forgetting, we construct a large-scale offline multimodal database—enabling continual adaptation through retrieval of similar cases during training. We conduct a comprehensive validation of our proposed framework against existing state-of-the-art methods; extensive experiments across 11 medical datasets, covering 9 modalities and 10 organs, demonstrate that our approach achieves state-of-the-art accuracy.

Keywords: Vision-Language Models · Continual Learning · Retrieval-Augmented Generation.

1 Introduction

The rapid evolution of Multimodal Vision-Language Models (VLMs) has sparked a paradigm shift in automated medical diagnosis, offering the potential to process complex clinical data akin to human experts. However, the deployment of these generalist models in real-world clinical settings faces a formidable obstacle: Continual Learning (CL) [30]. Unlike static benchmarks, clinical environments are dynamic, requiring models to incrementally acquire new diagnostic capabilities (e.g., identifying new diseases or adapting to new modalities) without forgetting previously learned knowledge [18,24]. While CL has been extensively studied in natural imaging, its application in the biomedical domain confronts a unique

and severe core dilemma: the distinct data distribution characteristics of medical imagery characterized by high intra-domain homogeneity and pronounced inter-domain heterogeneity (as shown in Fig. 1).

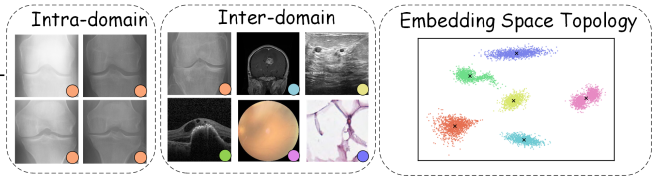

Fig. 1. Medical images are intra-domain uniform but cross-domain distinct.

This distributional property leads to two critical failures in standard CL frameworks: fine-grained feature confusion and blurred decision boundaries. In medical diagnostics, different pathological conditions—such as a benign adenomatous polyp and a malignant adenocarcinoma—may share highly similar visual features, distinguished only by subtle histological nuances like desmoplastic reaction [22]. When a VLM updates its parameters to learn a new task, the delicate decision boundaries established for these fine-grained distinctions are often the first to be eroded, leading to catastrophic forgetting. Existing approaches, such as retrieval-based replay or standard Retrieval-Augmented Generation (RAG) [13], attempt to mitigate forgetting by providing the model with relevant historical examples. However, these methods predominantly focus on retrieving positive semantic matches—reinforcing what a disease is. They critically overlook hard negatives or mimics—what a disease is not [25]. By failing to explicitly teach the model to distinguish between a lesion and its visual mimics, current systems remain prone to generating erroneous diagnoses based on superficial visual similarities, posing significant safety risks in clinical decision-making.

Experienced clinicians typically formulate diagnoses by synthesizing their clinical expertise with retrospective lessons from past misdiagnoses; their process relies not merely on observational pattern recognition but also on the rigorous exclusion of potential mimics. Drawing inspiration from this cognitive workflow, we propose a novel Misdiagnosis-Aware Structured Memory framework to address this challenge, empowering the model to perform diagnostics akin to a human expert. Diverging from conventional graph structures that solely map entities to their standard definitions, we construct a Hierarchical Pitfall Graph. This graph explicitly embeds misdiagnosis lessons into the retrieval topology by linking medical entities to their common mimics and specific exclusionary criteria (e.g., distinguish A from B by verifying feature X). This structured memory is integrated into a Retrieval-Augmented Generation (RAG) system, providing real-time, on-demand knowledge guidance throughout the model fine-tuning process. Concurrently, we construct a large-scale multimodal knowledge base designed to mirror the retrospective experience of human physicians. By revisit-

ing historical exemplars during the training process, this mechanism facilitates continual learning.

Our main contributions are summarized as follows:

1. To address the challenge of fine-grained feature confusion in medical imagery, we design a Misdiagnosis-Aware Hierarchical Pitfall Graph. By linking medical entities to their mimics and exclusionary criteria, this structure encodes diagnostic logic, thereby endowing the model with diagnostic capabilities akin to a clinician.
2. To address catastrophic forgetting, we construct a large-scale offline multimodal database. Serving as an external knowledge base analogous to a physician's past experience, the system facilitates continual adaptation by continuously retrieving similar cases during the training process.
3. We conduct a comprehensive validation across 11 medical datasets covering 9 modalities and 10 organs. The results demonstrate that our approach achieves state-of-the-art accuracy, significantly outperforming existing methods in distinguishing complex, visually similar pathologies.

2 Method

2.1 Problem Definition

We adopt the standard framework of generalist medical task-incremental learning, defined as a stream of T sequential tasks $\mathcal{S} = \{\mathcal{T}^1, \mathcal{T}^2, \ldots, \mathcal{T}^T\}$. Each task $\mathcal{T}^t$ is associated with a dataset $\mathcal{D}^t = \{(x_i, y_i)\}_{i=1}^{N_t}$, where $x_i \in \mathcal{X}$ denotes a medical image and $y_i \in \mathcal{C}^t$ represents its ground-truth label within the task-specific label space. Our objective is to train a multimodal foundation model f_θ, parameterized by θ and composed of image and text encoders ($E_I(\cdot)$ and $E_T(\cdot)$), to simultaneously satisfy two competing requirements at any time step t: plasticity, which entails efficiently learning new classes in $\mathcal{C}^t$ by minimizing prediction error on $\mathcal{D}^t$, and stability, which mandates retaining discriminative performance across all previously encountered label spaces $\bigcup_{k=1}^{t-1} \mathcal{C}^k$ without accessing full historical datasets $\mathcal{D}^{1:t-1}$.

2.2 Construction of the Hierarchical Pitfall Graph

To explicitly address the fine-grained misdiagnosis dilemma, we construct a Hierarchical Pitfall Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that synthesizes structured medical ontologies (UMLS [4], MeSH [3]) with unstructured clinical evidence mined from PubMed [6]. Unlike conventional knowledge graphs that rely solely on semantic hierarchy, our topology introduces novel adversarial connections to encode diagnostic logic: Mimicry Edges that link target entities to their visual mimics (e.g., Colon Adenocarcinoma $\leftrightarrow$ Pseudoinvasion), and Exclusionary Edges that map these mimics to specific absent attributes. As exemplified in Fig. 2, to resolve the confusion between malignant adenocarcinoma and benign pseudoinvasion—both characterized by deep glandular structures—the graph explicitly links the Pseudoinvasion

node to the attribute Desmoplastic Reaction via a "Lacks" relation, while the Adenocarcinoma node retains an "Exhibits" connection. This structure transforms static definitions into dynamic misdiagnosis lessons, enabling the system to retrieve structured exclusionary criteria that enforce discriminative margins during model fine-tuning.

2.3 Retrieval-Guided Continual Training

To address the stability-plasticity dilemma in medical continual learning, we introduce a Retrieval-Guided Continual Training framework. Unlike conventional continual learning methods that maintain a static, limited-capacity buffer, our approach dynamically constructs a knowledge replay stream by querying a large-scale offline multimodal database $\mathcal{B} = \{(x_i, t_i)\}_{i=1}^{M}$. This process ensures that the model's update on the current task $\mathcal{T}^t$ is constrained by a rich, semantically relevant historical context retrieved from the database, thereby aligning the decision boundaries of the current model θ^t with those of the frozen historical model θ^{t-1}.

Dense Retrieval and Visual-Semantic Reranking. We construct our large-scale offline multimodal database $\mathcal{B}$ using the Open-PMC-18M dataset [2], which comprises over 18 million biomedical image-text pairs derived from PubMed Central. We leverage a large-scale external database as a non-parametric memory bank, from which we dynamically retrieve historical exemplars to form a knowledge replay stream, thereby bypassing the limitations of static rehearsal buffers. Given an input image x from the current task, the system first generates a misdiagnosis lesson query, denoted as q_{rag}, derived from the aforementioned Hierarchical Pitfall Graph. This query explicitly encodes the discriminative features distinguishing the target pathology from its mimics. We first query the database using q_{rag} to retrieve a broad set of candidates based on text embedding similarity [19]. To filter out visually disparate outliers that match the text but not the visual features, we compute a comprehensive similarity score $S(x, x_k)$ for each candidate pair (x_k, t_k) in the database:

$$S(x, x_k) = \underbrace{\text{sim}(E_T(q_{rag}), E_T(t_k))}_{\text{Semantic Relevance}} + \gamma \cdot \underbrace{\text{sim}(E_I(x), E_I(x_k))}_{\text{Visual Affinity}} \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ represents the cosine similarity, E_T and E_I denote the text and image encoders respectively, and γ is a hyperparameter balancing semantic and visual relevance.

Dual-Model Logit Distillation. To effectively utilize these retrieved samples for stability, we adopt a knowledge distillation strategy that treats the previous model θ^{t-1} as a fixed teacher and the current model θ^t as an adaptive student. The objective is to force the student to mimic the teacher's response distribution over the retrieved historical concepts. For each sample x_h in the replay

batch $\mathcal{B}_{replay}$, we forward it through both models. We compute the image-text matching logits relative to their corresponding text descriptions t_h:

$$z_h^{t-1} = \tau \cdot \text{sim}(E_I(x_h; \theta^{t-1}), E_T(t_h; \theta^{t-1})), \quad z_h^t = \tau \cdot \text{sim}(E_I(x_h; \theta^t), E_T(t_h; \theta^t)) \tag{2}$$

where τ is the temperature parameter. We then minimize the L_2 distance between these logit distributions to enforce consistency:

$$\mathcal{L}_{distill} = \frac{1}{k} \sum_{j=1}^{k} \| z_{h_j}^t - z_{h_j}^{t-1} \|_2^2 \tag{3}$$

Minimizing $\mathcal{L}_{distill}$ ensures that the current model retains the discriminative reasoning of the historical model.

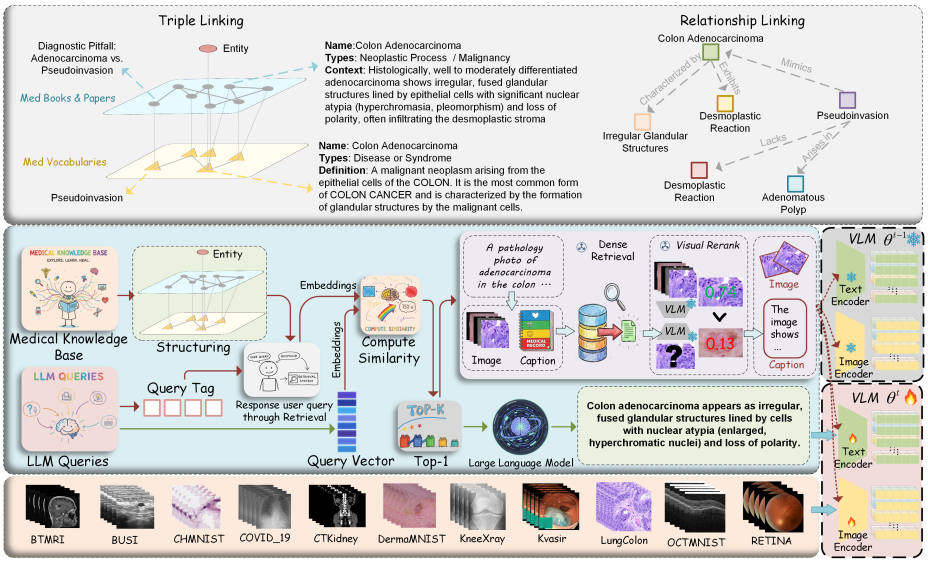

Fig. 2. Framework overview of our proposed method. Our framework utilizes a Hierarchical Pitfall Graph (HPG) to encode mimicry relationships. Inputs trigger a tag-based retrieval that aligns graph concepts with query embeddings to select the optimal discriminative prompt (Top-1) [27]. Subsequently, continual learning is enforced by distilling logits from historical exemplars, retrieved from an offline multimodal database via visual-semantic reranking.

Optimization Objective. The final training objective is a weighted combination of the task-specific plasticity loss and the retrieval-guided stability loss. For

Table 1. Comparison of SOTA methods on modality-wise order and random order. The best-performing method in each column is highlighted in bold. ACC (accuracy) and AUC (area under the ROC curve) measure classification precision and generalization capability, BWT (backward transfer) quantifies the retention of previously acquired knowledge after learning new tasks. The Δ denotes the performance differential relative to the Continual FT (continual fine-tuning).

Method	Publication	Modality-wise order						Random order					
		ACC	Δ	AUC.	Δ	BWT	Δ	ACC	Δ	AUC	Δ	BWT	Δ
Zero-shot	-	17.31	-	43.86	-	-	-	20.59	-	53.97	-	-	-
Continual FT	-	60.39	-	81.20	-	-23.64	-	56.18	-	79.45	-	-25.77	-
LwF [14]	*TPAMI 2017*	59.14	-1.25	80.08	-1.12	-11.20	+12.44	54.89	-1.29	77.65	-1.80	-17.22	+8.55
iCaRL [18]	*CVPR 2017*	62.37	+1.98	81.40	+0.20	-10.01	+13.63	60.92	+4.74	79.66	+0.21	-11.70	+14.07
WiSE-FT [25]	*CVPR 2022*	64.17	+3.78	84.55	+3.35	-5.62	+18.02	61.88	+5.70	80.16	+0.71	-9.32	+16.45
ZSCL [30]	*ICCV 2023*	56.12	-4.27	78.84	-2.36	-8.21	+15.43	47.69	-8.49	75.49	-3.96	-12.76	+13.01
MoE-CL [28]	*CVPR 2024*	65.13	+4.74	85.76	+4.56	-4.38	+19.26	62.14	+5.96	84.77	+5.32	-4.99	+20.78
SND [29]	*ECCV 2024*	62.73	+2.34	84.67	+3.47	-7.38	+16.26	60.03	+3.85	82.55	+3.10	-10.80	+14.97
DIKI [21]	*ECCV 2024*	66.99	+6.60	86.71	+5.51	-5.66	+17.98	62.47	+6.29	84.00	+4.55	-6.09	+19.68
GIFT [26]	*CVPR 2025*	67.27	+6.88	86.84	+5.64	-3.58	+20.06	66.50	+10.32	85.41	+5.96	-4.56	+21.21
CoOp [31]	*IJCV 2024*	62.14	+1.75	82.56	+1.36	-17.85	+5.79	61.32	+5.14	81.43	+1.98	-18.81	+6.96
BiomedCoOp [12]	*CVPR 2025*	67.02	+6.63	84.22	+3.02	-7.17	+16.47	65.33	+9.15	84.17	+4.72	-6.87	+18.90
Ours	-	**72.41**	**+12.02**	**89.37**	**+8.17**	**-2.44**	**+21.20**	**71.68**	**+15.50**	**88.79**	**+9.34**	**-3.10**	**+22.67**

a batch of current task data $\{(x, y)\}$, the total loss is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{InfoNCE}(x, y) + \lambda \cdot \mathcal{L}_{distill}(\mathcal{B}_{replay}) \tag{4}$$

here, $\mathcal{L}_{InfoNCE}$ optimizes the model to learn new distinctions in the current task, while the distillation term acts as a regularizer, anchoring the model's parameters to preserve its performance on the retrieved historical knowledge.

3 Experiments and Results

3.1 Datasets and Implementation Details

Our experimental validation is conducted on a diverse collection of 11 datasets that cover 9 different imaging modalities and 10 organ systems. This extensive setup includes Computerized Tomography (CTKidney [9]), Dermatoscopy (DermaMNIST [8,23]), Endoscopy (Kvasir [15]), Fundus Photography (RETINA [11,16]), Histopathology (LungColon [5], CHMNIST [10]), MRI (BTMRI [17]), Optical Coherence Tomography (OCTMNIST [10]), Ultrasound (BUSI [1]), and X-Ray imaging (COVID-19 [20], KneeXray [7]), thereby providing a robust assessment across complex biomedical contexts. All experiments were conducted on a workstation equipped with a 4-NVIDIA A6000 GPU, utilizing BiomedCLIP as the backbone and GPT-4 as the Large Language Model (LLM). We train each task for a fixed duration of 1,000 iterations, employing a learning rate of 1×10^{-5}, a label smoothing factor of 0.2, and a batch size of 64 per GPU. For datasets with limited samples, the training data was cycled to ensure the full iteration count was met.

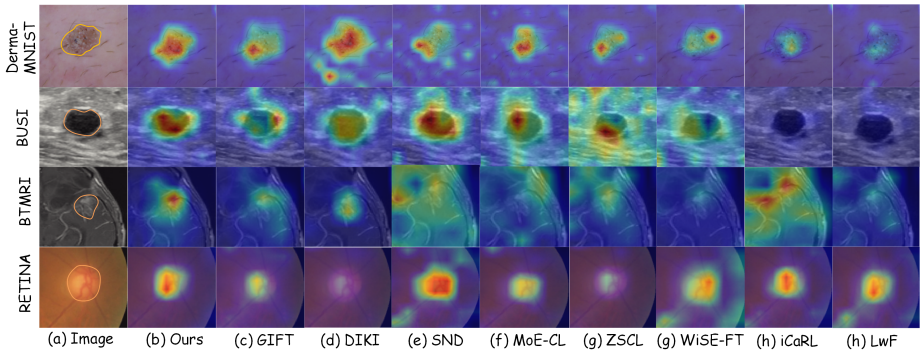

Fig. 3. Comparison of visual saliency maps generated by different methods.

3.2 Comparisons with the Previous Method

Table 1 presents the performance of different methods on 11 datasets under various task input orders. Notably, CoOp [31] and BiomedCoOp [12] are not inherently designed for continual learning. To ensure a fair comparison, we adapt these methods by applying a continual fine-tuning strategy. Under the most challenging Random order setting, our method achieves an ACC of 71.68% and an AUC of 88.79%. In comparison, the second-best method (GIFT [26]) attains an ACC of 66.50%.

Most methods suffer a substantial performance degradation when confronted with severe cross-domain shifts. For instance, the ACC of ZSCL [30] drops from 56.12% to 47.69%. In contrast, our method demonstrates strong robustness, with only marginal fluctuations in ACC (from 72.41% to 71.68%), while maintaining a relatively high BWT of -3.10%. Fig. 4 illustrates the accuracy of the first task under the random order training setting. This evidences that the proposed dynamic retrieval mechanism can effectively handle drastic inter-domain variations and stabilize the model's decision boundaries by retrieving relevant historical knowledge, regardless of the order in which tasks arrive. Fig. 3 presents a qualitative comparison of saliency maps produced by various approaches.

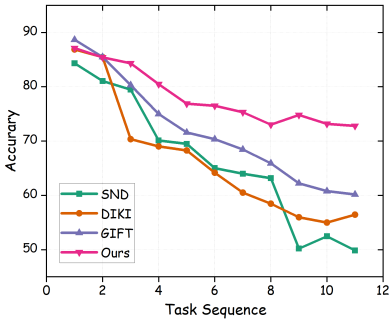

Fig. 4. ACC of 1st task in random order.

3.3 Ablation Study

Table 2 reports ablation results validating the contributions of each component. Specifically, removing the Hierarchical Pitfall Graph (w/o HPG), Visual-

Table 2. Ablation study is performed to evaluate the effectiveness of various components, including the construction of the hierarchical pitfall graph (HPG), visual-semantic reranking (VSR), dual-model logit distillation (DMLD). Δ denotes the performance gap compared to the full method (Ours (Full)).

	Method	Modality-wise order						Random order					
		ACC	Δ	AUC	Δ	BWT	Δ	ACC	Δ	AUC	Δ	BWT	Δ
Core Component Analysis	Ours (Base)	60.47	-11.94	80.51	-8.86	-19.19	-16.75	58.20	-13.48	79.63	-9.16	-20.48	-17.38
	w/o HPG	68.11	-4.30	83.29	-6.08	-8.39	-5.95	67.36	-4.32	82.19	-6.60	-8.79	-5.69
	w/o VSR	70.31	-2.10	87.10	-2.27	-4.19	-1.75	68.49	-3.19	86.50	-2.29	-4.54	-1.44
	w/o DMLD	66.14	-6.27	84.28	-5.09	-9.45	-7.01	65.98	-5.70	83.21	-5.58	-11.02	-7.92
	Ours (Full)	**72.41**	-	**89.37**	-	**-2.44**	-	**71.68**	-	**88.79**	-	**-3.10**	-
Open-Source LLM Comparison	GLM-4-Plus	69.30	-	85.89	-	-5.11	-	69.03	-	81.16	-	-7.09	-
	Qwen-3-32B	70.43	-	85.11	-	-5.58	-	69.71	-	84.10	-	-6.90	-
	DeepSeek-V3	71.48	-	88.05	-	-7.32	-	70.52	-	87.94	-	-7.97	-
	GPT-4.1-mini	71.59	-	87.99	-	-4.04	-	70.89	-	86.10	-	-4.77	-
Backbones Comparison	MMKD	65.89	-	79.98	-	-7.66	-	64.13	-	77.60	-	-7.73	-
	UniMed	70.12	-	86.41	-	-5.11	-	68.08	-	85.00	-	-5.87	-
	CLIP	69.53	-	85.44	-	-4.38	-	69.07	-	84.54	-	-5.02	-

Semantic Reranking (w/o VSR), or Dual-Model Logit Distillation (w/o DMLD) leads to notable drops in ACC and BWT, highlighting their roles in structured knowledge modeling and forgetting mitigation. Furthermore, we report experimental results obtained by substituting the core LLM and visual backbone within our framework with other state-of-the-art models. Fig. 5 provides qualitative visualization of retrieval results.

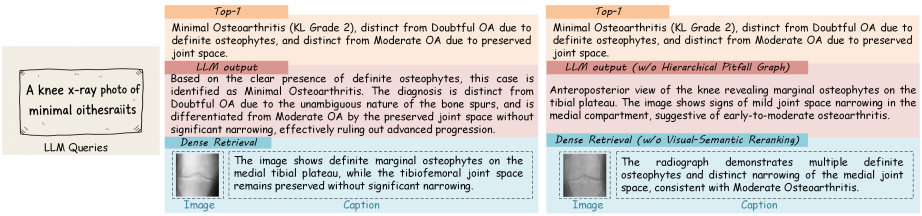

Fig. 5. In response to the query for a knee X-ray showing minimal osteoarthritis, our framework accurately retrieves a Grade 2 case with preserved joint space and generates a diagnosis incorporating explicit exclusionary logic. In contrast, removing Visual-Semantic Reranking leads to fine-grained feature confusion, resulting in the retrieval of a Moderate Osteoarthritis case. Furthermore, removing the Hierarchical Pitfall Graph causes the LLM to hallucinate mild joint space narrowing and yield an ambiguous output, failing to establish a clear decision boundary.

4 Discussion and Conclusion

In this study, we propose a Misdiagnosis-Aware Structured Memory framework to address fine-grained feature confusion in medical continual learning. By encod-

ing clinical pitfalls into a Hierarchical Pitfall Graph, our method converts static retrieval into dynamic, error-aware reasoning that sharpens decision boundaries. This process mirrors clinical reasoning by recalling discriminative knowledge before cross-referencing retrieved exemplars to enable continual adaptation. While this alignment with clinical cognition introduces architectural complexity, it is indispensable for resolving fine-grained visual nuances, as it explicitly incorporates exclusionary reasoning. Extensive validation across 11 datasets demonstrates consistent superiority over state-of-the-art (SOTA) methods, confirming that integrating structured negative knowledge is essential for mitigating catastrophic forgetting.

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

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
2. Baghbanzadeh, N., Islam, M.S., Ashkezari, S., Dolatabadi, E., Afkanpour, A.: Open-pmc-18m: A high-fidelity large scale medical dataset for multimodal representation learning. arXiv preprint arXiv:2506.02738 (2025)
3. Baumann, N.: How to use the medical subject headings (mesh). International journal of clinical practice **70**(2), 171–174 (2016)
4. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
5. Borkowski, A.A., Bui, M.M., Thomas, L.B., Wilson, C.P., DeLand, L.A., Mastorides, S.M.: Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142 (2019)
6. Canese, K., Weis, S.: Pubmed: the bibliographic database. The NCBI handbook **2**(1), 2013 (2013)
7. Chen, P.: Knee osteoarthritis severity grading dataset. Mendeley Data **1**(10.17632), 30784984 (2018)
8. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
9. Islam, M.N., Hasan, M., Hossain, M.K., Alam, M.G.R., Uddin, M.Z., Soylu, A.: Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from ct-radiography. Scientific Reports **12**(1), 11440 (2022)
10. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
11. Köhler, T., Budai, A., Kraus, M.F., Odstrčilik, J., Michelson, G., Hornegger, J.: Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In: Proceedings of the 26th IEEE international symposium on computer-based medical systems. pp. 95–100. IEEE (2013)

12. Koleilat, T., Asgariandehkordi, H., Rivaz, H., Xiao, Y.: Biomedcoop: Learning to prompt for biomedical vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14766–14776 (2025)
13. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
14. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
15. Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.T., Lux, M., Schmidt, P.T., et al.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference. pp. 164–169 (2017)
16. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)
17. Rahman, T., Islam, M.S.: Mri brain tumor detection and classification using parallel deep convolutional neural networks. Measurement: Sensors **26**, 100694 (2023)
18. 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)
19. Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Foundations and trends® in information retrieval **3**(4), 333–389 (2009)
20. Tahir, A.M., Chowdhury, M.E., Khandakar, A., Rahman, T., Qiblawey, Y., Khurshid, U., Kiranyaz, S., Ibtehaz, N., Rahman, M.S., Al-Maadeed, S., et al.: Covid-19 infection localization and severity grading from chest x-ray images. Computers in biology and medicine **139**, 105002 (2021)
21. Tang, L., Tian, Z., Li, K., He, C., Zhou, H., Zhao, H., Li, X., Jia, J.: Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models. In: European conference on computer vision. pp. 346–365. Springer (2024)
22. Tong, Y., Yuan, J., Zhang, M., Zhu, D., Zhang, K., Wu, F., Kuang, K.: Quantitatively measuring and contrastively exploring heterogeneity for domain generalization. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 2189–2200 (2023)
23. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
24. Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.Y., Ren, X., Su, G., Perot, V., Dy, J., et al.: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: European conference on computer vision. pp. 631–648. Springer (2022)
25. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7959–7971 (2022)
26. Wu, B., Shi, W., Wang, J., Ye, M.: Synthetic data is an elegant gift for continual vision-language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2813–2823 (2025)

27. Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., Grau, V.: Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. arXiv preprint arXiv:2408.04187 (2024)
28. Yu, J., Zhuge, Y., Zhang, L., Hu, P., Wang, D., Lu, H., He, Y.: Boosting continual learning of vision-language models via mixture-of-experts adapters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23219–23230 (2024)
29. Yu, Y.C., Chen, C.P.H.J.J., Chang, K.P., Lai, Y.H., Yang, F.E., Wang, Y.C.F.: Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models supplementary material
30. 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)
31. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)