Forging a Dynamic Memory: Retrieval-Guided Continual Learning for Generalist Medical Foundation Models

Supplementary Material

Appendix Contents

A. Limitations

Storage Usage. As with all Retrieval-Augmented Generation (RAG) methodologies, the construction and maintenance of our multimodal retrieval corpus entail additional computational resources and human effort. Furthermore, dynamic retrieval inherently imposes computational and temporal overheads. These challenges motivate us to advocate for a phased implementation of our approach. Given that the initial content of the Question Pool is static, retrieval for this segment can be pre-computed and reused. Consequently, we restrict real-time retrieval operations solely to questions added subsequently.

The Trade-off in Data Sources. To strictly prevent data leakage and avoid copyright or ethical disputes, we utilized the PubMed scientific literature database as our retrieval corpus. This choice ensures the reliability of our experimental results, the generalizability of the method in real-world deployments, and the overall safety and legality of the model. However, this imposes a limitation: application-level principles, such as fairness and broad ethical considerations, cannot be explicitly enforced at the algorithmic level during retrieval. Since the vast majority of scientific captions do not contain sensitive attributes like gender or race, and inferring such information solely from visual data is challenging, our continual learning method cannot explicitly target these dimensions. Instead, it primarily ensures the absence of fundamental safety hazards.

B. Ethical Statement

This research strictly adheres to the relevant ethical guidelines for medical AI research.

Data Usage and Patient Privacy. All data used in this study were sourced from publicly available research publications or scientific datasets. All data were fully anonymized and de-identified by the original providers prior to release and contain no Protected Health Information (PHI). Our usage strictly complies with the Data Use Agreement (DUA) for PubMed (consistent with BiomedCLIP [29]/BIOMEDICA [14]) and adheres to the DUAs of all respective open-source classification datasets involved.

Algorithmic Bias. The performance of our model is dependent upon both the backbone and the continual learning methodology. The data for both components are sourced from scientific literature available on PubMed. Although it is generally assumed that a dataset comprising tens of millions of samples provides sufficient data diversity, it cannot be guaranteed that undiscovered biases (e.g., in demographic representation across race, age, or sex) are not present. These biases may subsequently be learned and amplified by the model. Future work is required to specifically quantify and mitigate such biases.

C. Future Work

We will focus on the real-world rollout of PRIMED and making it easier to deploy.

Data Services. Acknowledging the difficulties associated with large-scale data management, we will release our retrieval database subject to a secondary ethical audit. We will also establish a cloud service where users can retrieve information by merely uploading questions. Additionally, we support data streaming to streamline local deployment and the augmentation of proprietary retrieval repositories.

Rare Disease Content Enrichment. To mitigate the data scarcity and heterogeneity of rare diseases, we plan to augment both the retrieval corpus and the Question Pool, thereby extending our framework's generalizability and value in this challenging domain.

D. Method and Evaluation Details

This section details the data cleaning methodologies employed to complement the construction of the retrieval corpus. It further elucidates the logic governing the curation of the benchmark dataset and offers comprehensive supplementary information. Lastly, we provide the disaggregated performance metrics for HieraMedTransfer, noting that the aggregated means of these values constitute the results presented in the main body of the paper.

D.1. Data Cleansing Approaches

Data Acquisition. We adopted the data acquisition methodology outlined in BIOMEDICA [14] to collect raw image-caption pairs. Since BIOMEDICA has already implemented fine-grained unsupervised clustering based on DINOv2 [16], we were able to exclude clinically irrelevant content—such as charts and natural images—by simply filtering based on the off-the-shelf pseudo-labels.

Multi-Subgraph Decoupling. Scientific literature frequently employs multi-panel figures to serve its illustrative purposes. However, given that clinical diagnostic images are predominantly presented in a single-panel format, the abundance of multi-panel content in reference datasets is detrimental to Continual Learning.

Drawing inspiration from the work [3] of Baghbanzadeh et al., we recognize that utilizing object detection models for multi-subgraph partitioning presents a promising approach. Given that subfigures in scientific literature typically consist of content with similar domains or semantics, we utilized single-panel images from our previously collected dataset for synthesis. Specifically, we combined images sharing the same coarse-grained pseudo-labels to generate synthetic multi-panel figures, resulting in a batched training dataset with object detection annotations. As depicted in Fig.1, we present a Multi-Subgraph Capture Model tailored for the medical domain, leveraging the DAB-DETR [13] architecture.

Ultimately, regular expressions were utilized to identify subfigure indicators (e.g., (1), (a), A). This facilitated the segmentation and realignment of captions based on spatial layout, given that scientific literature typically follows a fixed left-to-right reading sequence.

D.2. Benchamrk Database Construction

We constructed the MGTIL benchmark based on three key principles:

- There are substantial discrepancies across domains, which vary significantly in terms of imaging modalities and spatial resolutions.
- The dataset preserves fine-grained intra-domain variations, such as diverse spatial regions and varying object scales, thereby avoiding severe homogeneity.

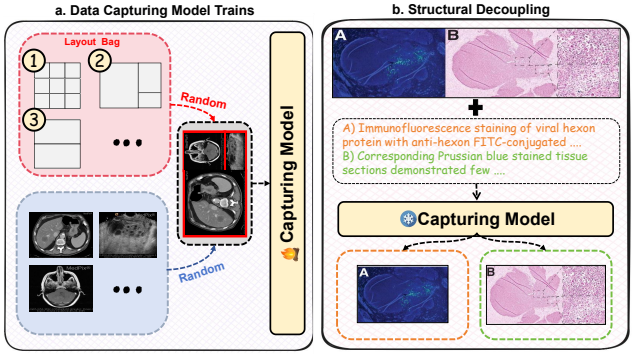


Figure 1. Overview of the Multi-Subgraph Capture Model: Training Methodology and Application Scenarios.

- These challenging datasets typify the common hurdles in medical classification tasks, being characterized by a large number of categories, high task complexity, and the prevalence of few-shot scenarios.

HieraMedTransfer Construction. As detailed in Tab. 1, our experimental evaluation encompasses three distinct modalities: X-ray, pathological, and fundus images. The intra-domain datasets exhibit multi-level discrepancies, including variations in anatomical regions, resolutions, and label granularity, which effectively mirrors the complex data heterogeneity inherent to real-world medical scenario.

MedXtreme Construction. MedXtreme encompasses six medical datasets across distinct domains, characterized by large label spaces, high task complexity, and significant domain shifts. It is designed to evaluate a model's capacity for learning and memory retention on challenging tasks within a continual fine-tuning setting. Notably, the inclusion of a substantial number of few-shot classes effectively simulates the dilemma of diagnosing rare diseases in clinical practice. Further details on the dataset are provided in Tab.2.

Detailed Performance Analysis. Tab.3 and Tab.4 detail the results on HieraMedTransfer for Order I and II. Task-specific metrics follow the ZSCL [30] protocol, with averages listed in the "Average" column as in the main text. Notably, we supplement these tables with "Fine-tune" results—representing the theoretical upper bound achieved by fine-tuning solely on the target dataset. As shown, our method yields the highest average performance among state-of-the-art competitors and, most notably, performs comparably to the Fine-tune upper bound.

We visualize the accuracy evolution of selected tasks from two distinct orders in Fig.2 (a) and (b). In the context of Vision-Language Models (VLMs), an optimal Continual Learning strategy is characterized by a **"mirrored Z-shaped"** curve. This signifies the effective maintenance of zero-shot capabilities prior to task acquisition and the mitigation of forgetting subsequent to learning. As demonstrated, our method aligns closely with this ideal profile.

Table 1. Visualization of the nine datasets utilized in HieraMedTransfer. We implemented a design for multi-scale transfer across both in-domain and out-of-domain settings.

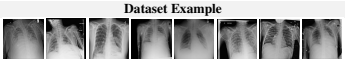
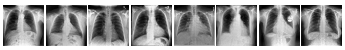
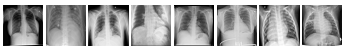
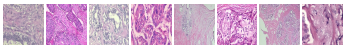
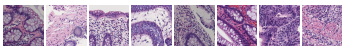
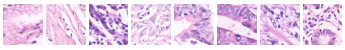
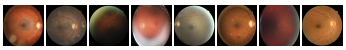
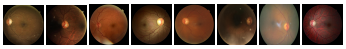
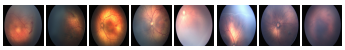
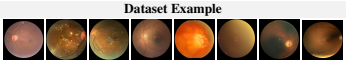
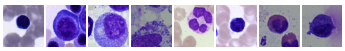
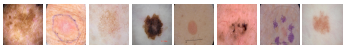
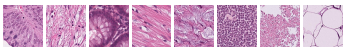
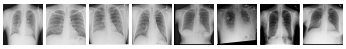
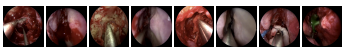
Dataset Example	Dataset Name	Domain	Region/Type	Number	Classes
	RANZCR [20]	X-ray	Blood Vessel	33665	11
	CheXchoNet [4]	X-ray	Chest	71589	4
	PD [2]	X-ray	Lung	4575	3
	Breakhis [21]	Patho.	Breast	7909	2
	Chaoyang [31]	Patho.	Colonic	6160	4
	Nucls [8]	Patho.	Gastric	33284	2
	Eyepacs [7]	Fundus	Diabetes	35126	5
	AIROGS [5]	Fundus	Glaucoma	101442	2
	FARFUM-RoP [1]	Fundus	ROP	1533	3

Table 2. Visualization of the six datasets utilized in MedXtreme. This collection was curated to maximize classification difficulty while satisfying the data requirements for fine-tuning.

Dataset Example	Dataset Name	Domain	Region/Type	Number	Classes
	AOD [19]	Fundus	Eye	10000	8
	BMC [15]	Cell	Bone Marrow	171375	21
	ISIC2024 [11]	Skin	Skin	81722	33
	NCT100K [9]	Patho.	Colorectal	100000	9
	NIH-Chest-Xray [24]	X-ray	Chest	112120	15
	PITVIS [6]	Endo.	Pituitary	120024	15

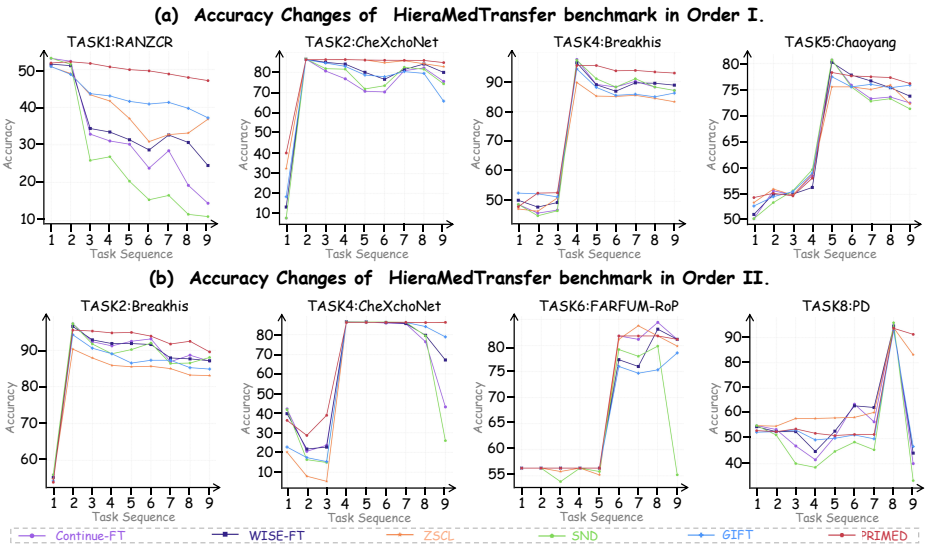


Figure 2. Illustration of classification accuracy changes as tasks are learned on the HieraMedTransfer benchmark in two orders. Our method consistently exhibits a mirrored Z-shaped pattern.

Table 3. Detailed Transfer, Avg., and Last scores (%) of different continue training methods on HieraMedTransfer benchmark in **Order I**. **Red Background** & **bold** indicate best results.

Method	RANZCR [20]	CheXchoNet [4]	PD [2]	Breakhis [21]	Chaoyang [31]	NuCLS [8]	Eyepacs [7]	AIROGS [5]	FARFUM-RoP [1]	Average
Zero-shot	16.44	28.34	52.07	53.88	52.65	57.45	62.28	96.46	56.49	52.90
Fine-tune	53.13	86.55	95.64	97.35	80.91	97.03	78.33	97.55	79.87	85.15
Transfer										
Continual FT		7.85	39.22	47.33	54.85	56.50	57.65	91.52	56.49	51.43
l_2 baseline		13.09	38.24	51.41	54.49	55.68	64.34	95.40	56.48	53.64
LwF [12]		31.89	43.68	**52.51**	54.13	56.49	19.96	34.30	56.49	43.68
iCaRL [18]		7.70	32.89	54.49	55.57	**60.60**	58.26	96.67	56.41	52.82
WiSE-FT [25]		13.33	40.85	49.31	54.37	55.85	60.66	94.51	56.49	53.17
ZSCL [30]		34.88	47.17	51.08	54.41	57.26	**64.43**	95.68	56.49	57.68
MoE-CL [27]		32.61	51.36	50.14	53.96	56.40	61.72	92.87	55.38	56.81
SND [28]		7.85	40.20	46.95	54.77	58.21	58.65	92.98	56.41	52.00
DIKI [22]		21.87	47.51	51.03	54.96	58.36	63.82	96.65	56.48	56.34
GIFT [26]		15.42	42.57	52.21	55.46	52.27	59.25	93.77	56.49	53.43
PRIMED$_{uni}$		27.46	**52.40**	50.91	54.98	55.77	62.73	96.21	56.49	57.12
PRIMED$_{dyn}$		**39.09**	48.38	51.24	**55.58**	55.42	63.68	**96.72**	**56.49**	**58.33**
Avg.										
Continual FT	31.69	70.13	80.90	75.81	66.16	72.41	59.83	91.84	59.38	67.57
l_2 baseline	27.71	76.31	77.25	76.05	66.25	71.94	66.82	95.35	58.94	68.51
LwF [12]	8.55	67.76	68.70	68.84	60.52	70.07	37.36	47.07	59.45	54.26
iCaRL [18]	49.69	77.56	71.53	76.41	64.92	**76.00**	64.42	**96.78**	59.09	70.71
WiSE-FT [25]	35.33	74.52	81.38	76.46	66.83	71.93	63.60	94.79	59.45	69.37
ZSCL [30]	43.91	78.98	80.39	72.74	64.74	70.76	67.67	95.90	59.09	70.46
MoE-CL [27]	33.07	80.58	**84.92**	77.92	65.58	70.87	67.40	96.84	59.43	70.73
SND [28]	25.77	71.23	82.21	75.96	65.88	73.43	60.05	93.50	59.02	67.45
DIKI [22]	41.84	79.50	80.74	75.70	64.81	72.58	67.61	96.28	59.41	70.94
GIFT [26]	42.97	72.68	80.37	75.73	66.92	71.39	63.57	95.95	58.29	69.76
PRIMED$_{uni}$	50.49	79.34	83.95	78.41	66.52	72.98	66.63	96.25	**59.45**	72.67
PRIMED$_{dyn}$	**50.95**	**80.83**	83.35	**79.82**	**67.71**	72.71	**67.73**	95.61	59.30	**73.11**
Last										
Continual FT	14.49	75.50	90.41	87.10	72.49	85.46	40.84	88.32	82.47	70.79
l_2 baseline	15.88	81.04	89.11	87.48	72.98	85.40	63.86	93.57	77.92	74.14
LwF [12]	1.92	75.67	70.59	72.06	64.40	67.56	64.31	86.23	83.12	65.10
iCaRL [18]	48.66	**86.23**	79.21	73.70	65.05	**94.62**	**75.93**	96.12	80.52	77.78
WiSE-FT [25]	24.51	79.94	90.63	88.87	73.79	86.42	55.48	94.28	83.12	75.23
ZSCL [30]	41.75	81.94	87.58	80.91	71.84	84.41	74.08	96.08	79.87	77.61
MoE-CL [27]	38.80	79.57	85.71	82.83	71.94	87.27	67.61	94.27	77.25	76.14
SND [28]	11.95	74.26	91.94	87.10	71.36	86.51	35.29	92.99	79.87	70.14
DIKI [22]	41.29	84.16	91.60	86.11	72.10	84.97	70.64	91.56	74.47	77.43
GIFT [26]	37.18	65.71	84.75	86.22	75.89	92.67	64.82	96.34	72.73	75.15
PRIMED$_{uni}$	**49.78**	84.77	91.94	91.66	74.60	93.30	70.64	95.96	**83.12**	81.75
PRIMED$_{dyn}$	47.07	84.80	**92.37**	**92.92**	**76.21**	92.31	74.74	**96.40**	81.82	**82.07**

Table 4. Detailed Transfer, Avg., and Last scores (%) of different continue training methods on HieraMedTransfer benchmark in **Order II**. **Red Background** & **bold** indicate best results.

Method	AIROGS [5]	Breakhis [21]	Chaoyang [31]	CheXchoNet [4]	Eyepacs [7]	FARFUM-RoP [1]	NuCLS [8]	PD [2]	RANZCR [20]	Average
Zero-shot	96.46	53.88	52.65	28.34	62.28	56.49	57.45	52.07	16.44	52.90
Fine-tune	97.55	97.35	80.91	86.55	78.33	79.87	97.03	95.64	53.13	85.15
Transfer										
Continual FT		56.01	56.15	28.83	64.18	56.49	55.10	52.63	11.98	47.67
l_2 baseline		55.12	55.82	18.58	64.80	56.62	54.15	53.52	7.40	45.75
LwF [12]		57.65	55.82	19.78	61.65	55.19	57.56	52.94	14.42	46.88
iCaRL [18]		53.73	54.69	38.39	**68.51**	56.49	**57.92**	48.80	6.57	48.14
WiSE-FT [25]		55.25	55.74	28.09	67.45	56.49	54.44	54.74	14.11	48.29
ZSCL [30]		55.25	55.10	9.42	67.86	51.43	54.31	56.83	11.36	45.20
MoE-CL [27]		52.87	55.50	28.78	63.12	55.93	53.29	53.09	18.25	47.60
SND [28]		**56.01**	56.40	24.34	64.42	55.84	56.04	46.31	15.96	46.92
DIKI [22]		52.06	55.02	20.54	64.76	55.63	54.77	49.19	20.57	46.57
GIFT [26]		54.36	54.86	18.51	67.72	56.49	55.06	51.36	15.95	46.79
PRIMED$_{uni}$		53.98	54.86	18.90	66.23	56.49	53.18	**56.92**	**22.62**	47.90
PRIMED$_{dyn}$		53.78	**56.40**	**34.66**	64.28	**56.89**	52.09	53.62	16.44	**48.52**
Avg.										
Continual FT	79.46	87.18	71.84	61.17	48.53	67.59	68.67	55.94	16.55	61.88
l_2 baseline	94.25	84.67	69.83	62.98	62.75	64.79	67.26	56.77	12.18	63.94
LwF [12]	89.64	72.64	63.22	53.50	63.82	65.87	70.31	59.04	17.45	61.72
iCaRL [18]	96.93	87.81	68.16	**70.34**	**73.58**	61.25	**70.69**	54.01	11.20	66.00
WiSE-FT [25]	92.30	86.91	72.14	64.70	57.51	66.66	66.65	57.98	18.37	64.80
ZSCL [30]	96.70	82.92	69.04	60.61	71.55	56.85	66.86	**64.66**	15.92	65.01
MoE-CL [27]	95.88	88.51	69.98	63.57	68.41	64.76	65.11	59.62	18.73	66.06
SND [28]	81.36	86.31	71.20	58.23	48.53	63.49	69.46	50.37	20.05	61.00
DIKI [22]	96.95	86.80	69.94	61.02	69.17	62.86	65.44	62.69	16.39	65.70
GIFT [26]	95.44	84.34	72.19	62.61	61.64	65.22	68.37	55.46	19.85	65.01
PRIMED$_{uni}$	96.51	88.12	72.08	62.21	69.54	66.74	67.33	60.78	25.81	67.68
PRIMED$_{dyn}$	**97.00**	**89.06**	**72.19**	69.06	69.34	**67.68**	66.46	61.24	**20.32**	**68.04**
Last										
Continual FT	39.17	86.98	71.20	43.22	17.29	81.17	95.46	40.09	**53.09**	58.63
l_2 baseline	78.62	86.47	66.02	81.80	38.91	66.23	91.59	41.18	50.45	66.81
LwF [12]	59.47	67.89	63.92	40.52	48.05	74.68	95.46	59.69	45.65	61.70
iCaRL [18]	96.79	84.07	64.72	86.23	**77.33**	62.99	94.82	49.67	48.26	73.88
WiSE-FT [25]	76.65	87.10	74.76	67.10	23.61	81.17	86.00	44.23	52.46	65.90
ZSCL [30]	96.67	84.96	70.71	86.13	74.51	45.45	91.17	89.98	52.33	76.88
MoE-CL [27]	81.34	85.95	70.27	76.56	64.75	71.58	91.47	75.99	50.02	74.21
SND [28]	47.02	87.99	72.33	26.00	9.71	55.19	95.55	33.33	52.76	53.32
DIKI [22]	90.07	85.97	73.11	78.35	68.66	75.32	90.84	80.31	48.03	76.74
GIFT [26]	90.59	84.83	75.08	78.87	41.24	78.57	94.83	46.84	51.04	71.32
PRIMED$_{uni}$	**96.90**	88.37	75.24	78.33	70.66	78.57	**96.40**	57.12	51.31	76.99
PRIMED$_{dyn}$	91.93	**89.51**	**75.73**	**86.31**	68.61	**81.17**	95.01	**91.29**	51.31	**81.21**

E. Ablation Study

To validate the effectiveness of our experimental settings at all levels, we performed extensive ablation studies involving four sequences on our two proposed benchmarks. The analysis is organized as follows: experimental setup, module-level ablation, hyperparameter and component-level ablation, and Dynamic Retrieval analysis.

E.1. Full Settings

To ensure reproducibility, we detail the key configurations and experimental settings for training our model as follows:

- **Batch Size and Label Smoothing:** We employ a batch size of 64 per GPU and apply label smoothing of 0.2. Notably, fine-tuning is fixed at 1,000 iterations across all datasets; for datasets with insufficient samples, the training data is cycled to meet this requirement.
- **Learning Rate:** A unified learning rate of 1×10^{-5} is applied across all regularization, replay, and distillation methods [12, 18, 25, 26, 28, 30]. For approaches based on LoRA [27] or Prompt Tuning [22], we strictly adhere to the hyperparameter settings outlined in their papers.
- **Detailed Configurations:** In the following sections, we present a comprehensive ablation study covering all relevant components and hyperparameters. For clarity, the default settings adopted in our method are highlighted with a **Red Background**.

E.2. Modular Level Analysis

As shown in Tab.5, we conducted module-level ablation studies across all benchmarks and sequences. Encouragingly, the results remain fully consistent with the conclusions presented in the main text. This demonstrates the exceptional robustness of our method and the synergistic coupling between modules.

E.3. Component and Hyperparameter Analysis

Tab.6 presents the ablation study demonstrating robustness at both the component and parameter levels, while consistently maintaining superior performance. Notably, under the challenging task scenarios simulated by MedXtreme, our retrieval method achieved a significantly larger performance margin compared to other approaches. This trend aligns with the behaviors observed in dynamic recall on uniformly distributed reference datasets. This suggests that in challenging scenarios or complex clinical settings, retrieval mechanisms with higher quality and finer granularity possess significant efficacy and potential.

E.4. Dynamic Retrieval Dataset Analysis

The Dynamic Retrieval component is the cornerstone of PRIMED and constitutes the fundamental difference between our method and existing approaches. Two specific aspects warrant further investigation. First, akin to other methods relying on reference datasets, the size of the dataset presents a critical trade-off. Insufficient capacity risks compromising generalization and diversity, while excessive size imposes a prohibitive computational and storage overhead. Since prior studies have demonstrated that performance tends to plateau beyond a certain threshold, our objective is to identify this optimal saturation point illustrated in Fig.3. Next, building on the determined peak number, we investigated the ratios for dynamic retrieval. Operating under the premise that task weights should exceed domain weights, which in turn should exceed general weights, we employed a grid search to identify the optimal ratios. The quantitative results regarding the dataset capacity saturation are detailed in Tab. 7, while the outcomes of the grid search for optimal retrieval ratios are tabulated in Tab. 8.

The experimental results highlight distinct requirements for recall versus generalization across different tasks. Specifically, for intra- and cross-domain transfer tasks such as HieraMedTransfer, enhanced generalization is pivotal for handling diverse scenarios. Conversely, in high-difficulty benchmarks like MedXtreme, models require an extensive, potentially iterative review of representative exemplars. This is an intriguing finding, as it parallels human cognitive processes that combine long-term retention with short-term intensive reinforcement. Indeed, many characteristics of model memory appear to mirror those inherent to human memory.

F. Backbone Selection and Generalizability

Prior to selecting the specific backbones, we briefly review the list of candidate models, providing a concise overview of these contrastive learning-based foundation models.

- **BiomedCLIP [29]** is a multimodal foundation model pretrained on PMC-15M, a large-scale dataset of 15 million image-text pairs sourced from 4 million scientific articles
- **MMKD-CLIP [23]** is a generalist biomedical foundation model developed via multi-teacher knowledge distillation, utilizing 19.2 million image-text feature pairs synthesized by 9 expert models from the PMC-OA dataset.
- **BMC-CLIP [14]** is trained on the large-scale BIOMEDICA dataset, which includes over 24M image-text pairs from over 6M open-access scientific articles.
- **UniMed-CLIP [10]:** is a unified vision-language model trained on UniMed, a large-scale open-source dataset of 5.3 million image-text pairs spanning six imaging modalities (X-ray, CT, MRI, Ultrasound, Pathology, Fundus).
- **CLIP [17]:** is a multimodal foundation model pretrained on WIT-400M, a large-scale dataset of 400 million image-text pairs collected from a variety of publicly available sources on the internet.

Table 5. Ablation study of different modules on HieraMedTransfer and MedXtreme. **Red Background** indicates the full model.

+CKT	+CMC	+DFG	HieraMedTransfer I			HieraMedTransfer II			MedXtreme I			MedXtreme II		
			Transfer	Avg.	Last	Transfer	Avg.	Last	ACC	AUC	BWT	ACC	AUC	BWT
✓			56.2	71.8	**82.6**	46.7	67.8	81.0	66.7	87.1	-4.2	65.9	86.3	-5.3
	✓		54.3	68.9	78.9	49.5	67.2	76.4	64.9	85.4	-7.6	60.4	83.7	-13.2
✓	✓		56.9	71.6	82.2	48.2	67.9	81.0	66.5	85.9	-4.8	64.9	85.7	-6.8
✓		✓	57.2	72.8	82.5	46.8	67.8	80.9	66.7	87.3	-4.3	65.8	86.1	-5.4
	✓	✓	56.7	70.2	77.1	**50.0**	67.4	77.2	65.2	85.2	-7.1	60.4	82.8	- 13.3
✓	✓	✓	**58.3**	**73.1**	82.1	48.5	**68.0**	**81.2**	**68.6**	**87.4**	**-2.7**	**68.1**	**86.3**	**-3.4**

Table 6. Ablation study on different components and hyperparameters. **Red Background** indicates optimal settings.

Comp./Hparam.		HieraMedTransfer I			HieraMedTransfer II			MedXtreme I			MedXtreme II		
Aspect	**Detail**	Transfer	Avg.	Last	Transfer	Avg.	Last	ACC	AUC	BWT	ACC	AUC	BWT
Teacher	Initial CLIP	57.9	70.7	78.2	47.7	66.9	80.1	60.5	80.7	-9.5	60.3	78.9	-11.2
	Last CLIP	**58.3**	**73.1**	**82.1**	**48.5**	**68.0**	**81.2**	**68.6**	**87.4**	**-2.7**	**68.1**	**86.3**	**-3.4**
	WISE(0.5)	58.1	71.4	79.1	48.3	67.1	80.5	63.5	83.1	-5.6	62.7	82.0	-8.1
KD Loss	Image-only	57.7	72.4	81.3	47.9	67.6	80.8	67.6	87.0	-3.9	67.0	86.2	-4.5
	Text-only	58.2	73.0	80.5	**49.8**	67.9	80.4	65.5	85.8	-6.3	63.2	85.3	-9.2
	Contra.	**58.3**	**73.1**	**82.1**	48.5	**68.0**	**81.2**	**68.6**	**87.4**	**-2.7**	**68.1**	**86.3**	**-3.4**
CKT Scale	$\alpha = 0.5$	58.3	72.8	81.3	**48.7**	67.9	80.6	66.5	86.9	-5.1	66.7	85.4	-4.4
	$\alpha = 1$	**58.3**	**73.1**	**82.1**	48.5	**68.0**	**81.2**	**68.6**	**87.4**	**-2.7**	**68.1**	**86.3**	**-3.4**
	$\alpha = 1.5$	58.3	73.0	81.9	48.0	67.8	81.0	66.2	87.4	-4.9	66.3	85.6	-5.1
CMC Scale	$\beta = 0.0$	57.2	72.8	**82.5**	46.8	67.8	80.9	66.7	87.3	-4.3	65.8	86.1	-5.4
	$\beta = 0.25$	**58.3**	**73.1**	82.1	**48.5**	**68.0**	**81.2**	**68.6**	**87.4**	**-2.7**	**68.1**	**86.3**	**-3.4**
	$\beta = 0.5$	58.1	72.5	81.4	48.2	67.4	77.3	66.8	85.7	-4.6	64.2	84.8	-7.9
Reg.	l_2	57.2	72.9	81.8	48.2	67.9	81.0	66.5	86.9	-4.8	64.9	86.4	-6.9
	EWC	57.2	71.3	82.1	48.0	66.8	**81.4**	67.3	87.2	-3.7	67.2	**86.9**	-3.5
	DFG	**58.3**	**73.1**	**82.1**	**48.5**	**68.0**	81.2	**68.6**	**87.4**	**-2.7**	**68.1**	86.3	**-3.4**
RAG	BM25	55.1	72.2	81.8	47.5	67.5	79.4	67.4	86.8	-3.4	64.8	85.5	-6.5
	Embedding	56.1	72.5	81.5	47.3	67.2	76.7	66.3	86.9	-4.8	63.9	85.4	-7.5
	Hierarchical	**58.3**	**73.1**	**82.1**	**48.5**	**68.0**	**81.2**	**68.6**	**87.4**	**-2.7**	**68.1**	**86.3**	**-3.4**

Table 7. Dynamic Retrieval Analysis in HieraMedTransfer.

Ratio	HieraMedTransfer I			HieraMedTransfer II		
$b:c$	Trans.	Avg.	Last	Trans.	Avg.	Last
10:90	58.0	72.8	81.7	48.3	67.8	80.8
20:80	58.3	73.0	82.0	48.2	67.9	81.1
30:80	**58.3**	**73.1**	**82.1**	**48.5**	**68.0**	**81.2**
40:80	58.3	73.0	82.0	48.4	67.8	79.8
50:50	58.3	73.0	81.7	48.5	68.0	80.2
60:40	58.2	72.9	82.1	48.0	67.9	80.8
80:20	58.2	73.0	81.4	48.4	68.0	81.2
90:10	58.2	73.1	81.5	48.1	67.6	78.6

Table 8. Dynamic Retrieval Analysis in MedXtreme.

Ratio	MedXtreme I			MedXtreme II		
$b:c$	ACC	AUC	BWT	ACC	AUC	BWT
100:100	68.4	87.0	**-2.4**	67.9	85.9	-3.5
0:100	68.2	86.5	-3.2	67.7	85.7	-3.9
70:70	68.6	86.9	-2.7	68.1	85.7	-3.4
70:90	68.4	87.0	-2.9	67.8	85.8	-3.5
70:100	**68.6**	**87.4**	-2.7	**68.1**	**86.3**	**-3.4**
50:100	68.5	87.3	-2.8	68.1	86.3	-3.4
50:75	68.3	87.1	-3.0	67.7	86.0	-4.2
90:100	68.2	86.9	-2.5	67.9	85.7	-3.7

F.1. Justification for the BiomedCLIP Backbone

Our choice of BiomedCLIP as the primary backbone is motivated by several factors, outlined below in descending order of significance. Crucially, we must underscore that this selection was not predicated solely on performance metrics. Indeed, considerations such as the guarantee against data contamination and the maturity of the architectural framework took precedence over raw performance.

Data Security. Admittedly, while all the aforementioned methods utilize open-source datasets, only BiomedCLIP and BMC-CLIP feature a comprehensive data acquisition

Table 9. Experimental Results of Continual Learning on Backbones Other than BiomedCLIP

Architecture		HieraMedTransfer I			HieraMedTransfer II			MedXtreme I			MedXtreme II		
Backbone	**Method**	Transfer	Avg.	Last	Transfer	Avg.	Last	ACC	AUC	BWT	ACC	AUC	BWT
MMKD [23]	Continual FT	45.8	66.9	79.4	45.7	66.8	79.5	65.7	87.5	-7.1	64.5	85.7	-8.6
	WiSE-FT [25]	46.4	67.1	80.2	46.4	67.1	80.0	64.8	87.3	-4.0	62.4	86.0	-6.9
	ZSCL [30]	**58.6**	69.3	77.8	44.8	65.0	79.8	57.4	81.4	-8.6	56.0	80.9	-10.4
	GIFT [26]	49.6	68.3	80.5	49.6	68.3	80.4	69.7	88.2	-1.8	68.4	88.1	-3.1
	PRIMED$_{dyn}$	50.7	**69.4**	**80.7**	**50.7**	**69.4**	**80.6**	**70.4**	**88.3**	**-1.5**	**70.5**	**88.1**	**-1.3**
UniMed [10]	Continual FT	45.0	66.4	79.8	44.9	66.4	80.0	61.0	83.4	-10.5	58.8	82.6	-13.7
	WiSE-FT [25]	44.0	66.5	81.5	44.0	66.5	81.4	57.9	83.2	-9.1	61.6	86.5	-5.4
	ZSCL [30]	44.9	66.5	81.4	42.5	63.8	81.9	63.2	85.0	-6.3	59.9	83.6	-10.2
	GIFT [26]	44.2	66.7	82.5	44.2	66.7	82.4	67.1	**88.7**	-3.6	67.0	87.6	-3.5
	PRIMED$_{dyn}$	**45.0**	**67.1**	**83.1**	**45.1**	**67.1**	**83.2**	**67.4**	88.4	-3.3	**67.2**	**88.3**	**-3.3**
CLIP [17]	Continual FT	34.9	59.1	71.1	34.2	58.6	54.4	61.5	83.7	-17.2	47.3	82.5	-34.1
	WiSE-FT [25]	34.8	60.3	70.7	34.9	61.6	72.6	64.2	85.1	-12.4	57.3	84.9	-20.8
	ZSCL [30]	35.7	61.1	78.0	35.5	61.0	78.5	63.0	82.3	-11.0	56.2	82.9	-19.2
	GIFT [26]	**36.2**	61.7	73.9	**36.2**	62.6	75.5	71.0	87.5	-6.0	70.9	88.7	-5.9
	PRIMED$_{dyn}$	35.7	**62.6**	**84.6**	35.6	**62.6**	**84.0**	**73.8**	**89.1**	**-2.6**	**73.7**	**89.4**	**-2.8**

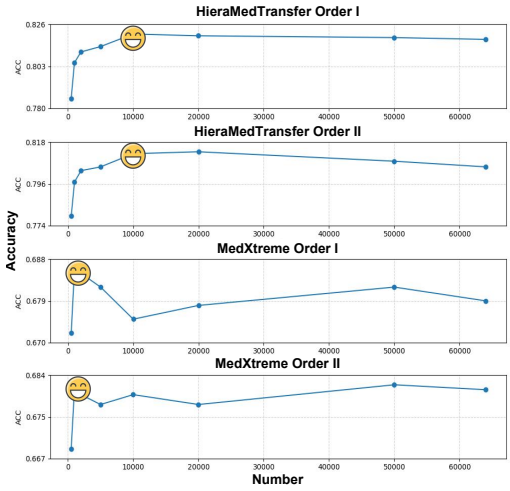


Figure 3. Peak Performance of Dynamic Retrieval across Datasets

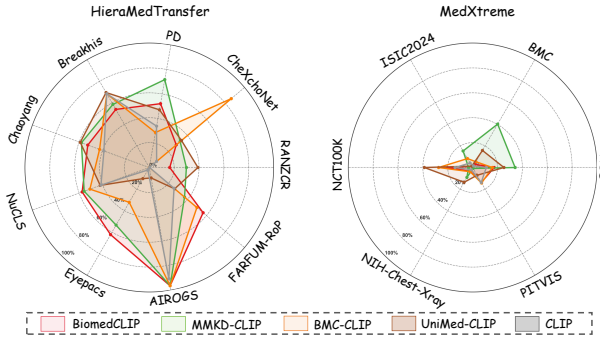


Figure 4. The zero-shot capabilities of 5 backbones across HieraMedTransfer and MedXtreme are depicted in radar chart format.

architecture. This distinction fundamentally ensures data integrity and reproducibility while preventing data contamination. Although MMKD-CLIP exhibits impressive experimental performance, it is derived from multi-model distillation, making it difficult to fully verify its data provenance. Therefore, establishing a more controllable baseline is of paramount importance to our work.

Zero-shot performance. Fig.4 presents the zero-shot results of various models on the HieraMedTransfer and MedXtreme benchmarks. For HieraMedTransfer, it is essential that the backbone exhibits a reasonable baseline of zero-shot capability; otherwise, the subsequent transferability evaluation would lack validity. BiomedCLIP demonstrates consistent performance across all datasets without exhibiting anomalous outlier peaks, establishing it as a highly robust and reliable candidate. In contrast, all models yield suboptimal performance on MedXtreme. Consequently, absolute performance metrics are of secondary importance compared to the potential risk of data contamination. In this regard, BiomedCLIP serves as a good choice.

Model Architecture. We favored a mature architecture that fits our specific demands; specifically, BiomedCLIP features a well-developed fine-tuning and post-training ecosystem. Additionally, we aimed to maximize experimental comparability by aligning with the ViT-B configuration used in natural image studies (e.g., ZSCL). Therefore, absent any distinct performance benefits, the ViT-L versions of BMC-CLIP and UniMed-CLIP were not selected as backbones for the main experiments.

F.2. Generalizable SOTA Performance

Although we consider BiomedCLIP to be the most intuitively suitable backbone, we also evaluated other ViT-B based backbones, as shown in Tab.9. Our method consistently achieved superior performance, demonstrating the effectiveness of our proposed strategy.

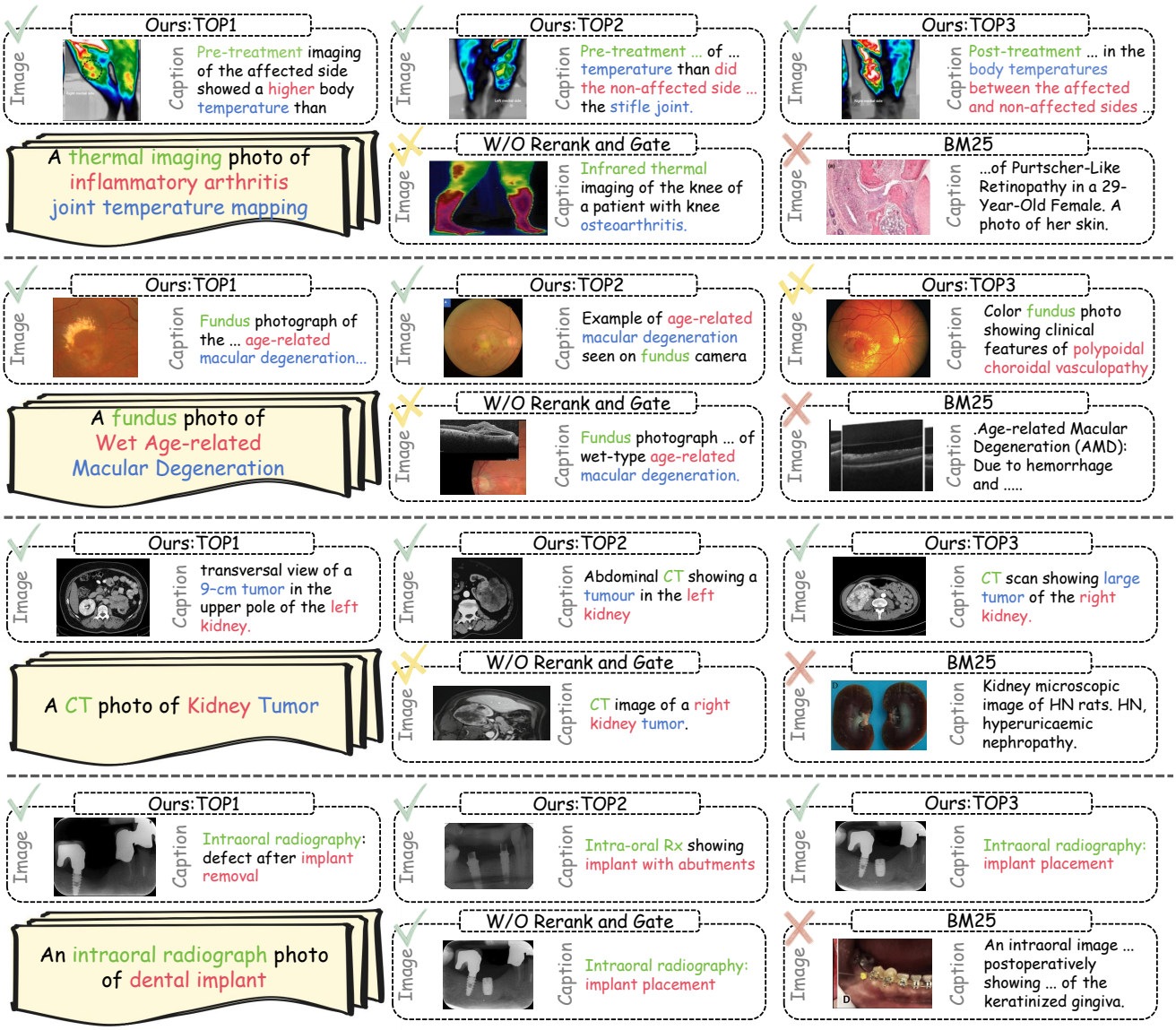


Figure 5. Qualitative comparison with state-of-the-art methods. Our method achieves superior performance in visual correction and textual precision. By explicitly aligning the hierarchical content within questions with the retrieved data, our method achieves optimal fine-grained retrieval performance. Furthermore, leveraging visual-level retrieval capabilities allows our approach to prioritize complete and high-quality images rather than relying solely on textual cues. This capability enhances the dynamic retrieval database's distillation.

G. Retrieval Visualization

Echoing the main text, we underscore the distinct superiority of our retrieval approach in terms of intuitive visualization. Primarily, the intrinsic mechanism of multimodal retrieval endows our method with strong semantic disentanglement capabilities. A prime example is in dentistry, where our model clearly discriminates between OCT scans and natural images, despite their high textual semantic overlap. Furthermore, our approach exhibits enhanced recall precision, moving beyond the rigid constraints of keyword matching. Since our data source relies heavily on multi-subgraph disentanglement, as detailed above, we effectively filter out cases of failed disentanglement or conceptual ambiguity—providing a robust guarantee of effectiveness. Ultimately, we are encouraged to observe that the retrieved content demonstrates both generalizability and hierarchical progression. By moving beyond isolated disease categories to account for the holistic connections between diseases, lesions, and subtypes, we believe this property is pivotal in improving model memorization.

References

[1] Morteza Akbari, Hamid-Reza Pourreza, Elias Khalili Pour, Afsar Dastjani Farahani, Fatemeh Bazvand, Nazanin Ebrahimiadib, Marjan Imani Fooladi, and Fereshteh Ramazani K. Farfum-rop, a dataset for computer-aided detection of retinopathy of prematurity. *Scientific Data*, 11(1): 1176, 2024. 3, 4, 5

[2] Amanullah Asraf and Zabirul Islam. Covid19, pneumonia and normal chest x-ray pa dataset. https://data.mendeley.com/datasets/jctsfj2sfn/1, 2021. Mendeley Data. 3, 4, 5

[3] Negin Baghbanzadeh, Sajad Ashkezari, Elham Dolatabadi, and Arash Afkanpour. Open-pmc-18m: A high-fidelity large scale medical dataset for multimodal representation learning. *arXiv preprint arXiv:2506.02738*, 2025. 2

[4] Shreyas Bhave, Victor Rodriguez, Timothy Poterucha, Simukayi Mutasa, Dwight Aberle, Kathleen M Capaccione, Yibo Chen, Belinda Dsouza, Shifali Dumeer, Jonathan Goldstein, et al. Deep learning to detect left ventricular structural abnormalities in chest x-rays. *European heart journal*, 45 (22):2002–2012, 2024. 3, 4, 5

[5] Coen De Vente, Koenraad A Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, Daniel Truhn, Temirgali Aimyshev, Yerkebulan Zhanibekuly, Tien-Dung Le, et al. Airogs: Artificial intelligence for robust glaucoma screening challenge. *IEEE transactions on medical imaging*, 43 (1):542–557, 2023. 3, 4, 5

[6] EndoVis. Endoscopic vision challenge 2023. https://doi.org/10.5281/zenodo.8315050, 2023. Zenodo. 3

[7] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016. 3, 4, 5

[8] Weiming Hu, Chen Li, Xiaoyan Li, Md Mamunur Rahaman, Jiquan Ma, Yong Zhang, Haoyuan Chen, Wanli Liu, Changhao Sun, Yudong Yao, et al. Gashissdb: A new gastric histopathology image dataset for computer aided diagnosis of gastric cancer. *Computers in biology and medicine*, 142: 105207, 2022. 3, 4, 5

[9] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. https://doi.org/10.5281/zenodo.1214456, 2018. Zenodo. 3

[10] Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities. *arXiv preprint arXiv:2412.10372*, 2024. 6, 8

[11] Nicholas Kurtansky, Veronica Rotemberg, Maura Gillis, Kivanc Kose, Walter Reade, and Ashley Chow. Isic 2024 - skin cancer detection with 3d-tbp. https://kaggle.com/competitions/isic-2024-challenge, 2024. Kaggle. 3

[12] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 4, 5, 6

[13] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 2

[14] Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Anita Rau, Austin Wolfgang Katzer, et al. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19724–19735, 2025. 1, 2, 6

[15] Christian Matek, Sebastian Krappe, Christian Münzenmayer, Torsten Haferlach, and Carsten Marr. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood, The Journal of the American Society of Hematology*, 138(20):1917–1927, 2021. 3

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 6, 8

[18] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 4, 5, 6

[19] Mohammad Riadur Rashid, Shayla Sharmin, Tania Khatun, Md Zahid Hasan, and Mohammad Shorif Uddin. Eye disease image dataset. https://data.mendeley.com/datasets/s9bfhswzjb/1, 2024. Mendeley Data. 3

[20] Jarrel Seah, Jen, Maggie, Meng Law, Phil Culliton, and Sarah Dowd. Ranzcr clip - catheter and line position challenge. https://kaggle.com/competitions/ranzcr-clip-catheter-line-classification, 2020. Kaggle. 3, 4, 5

[21] Fabio A Spanhol, Luiz S Oliveira, Caroline Petitjean, and Laurent Heutte. A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7):1455–1462, 2015. 3, 4, 5

[22] Longxiang Tang, Zhuotao Tian, Kai Li, Chunming He, Hantao Zhou, Hengshuang Zhao, Xiu Li, and Jiaya Jia. Mind the interference: Retaining pre-trained knowledge in parameter efficient continual learning of vision-language models. In *European conference on computer vision*, pages 346–365. Springer, 2024. 4, 5, 6

[23] Shansong Wang, Zhecheng Jin, Mingzhe Hu, Mojtaba Safari, Feng Zhao, Chih-Wei Chang, Richard LJ Qiu, Justin

Roper, David S Yu, and Xiaofeng Yang. Unifying biomedical vision-language expertise: Towards a generalist foundation model via multi-clip knowledge distillation. *arXiv preprint arXiv:2506.22567*, 2025. 6, 8

[24] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 3

[25] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 4, 5, 6, 8

[26] Bin Wu, Wuxuan Shi, Jinqiao Wang, and Mang Ye. Synthetic data is an elegant gift for continual vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2813–2823, 2025. 4, 5, 6, 8

[27] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. 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*, pages 23219–23230, 2024. 4, 5, 6

[28] Yu-Chu Yu, Chi-Pin Huang, Jr-Jen Chen, Kai-Po Chang, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models. In *European Conference on Computer Vision*, pages 219–236. Springer, 2024. 4, 5, 6

[29] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 1, 6

[30] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19125–19136, 2023. 2, 4, 5, 6, 8

[31] Chuang Zhu, Wenkai Chen, Ting Peng, Ying Wang, and Mulan Jin. Hard sample aware noise robust learning for histopathology image classification. *IEEE transactions on medical imaging*, 41(4):881–894, 2021. 3, 4, 5