CoLoRA-Med: Continual Medical Learning with Low-Rank Adaptation

Anonymous authors
Paper under double-blind review

Abstract

Vision-language models must continually adapt to new tasks without catastrophic forgetting of earlier ones, a requirement that is especially critical in medical applications, where reliability depends on retaining previously acquired capabilities. Existing low-rank adaptation methods for continual learning either constrain parameter updates or isolate them per task, and thus struggle to simultaneously share knowledge across related tasks and prevent interference between conflicting ones. We introduce CoLoRA-Med, a continual learning method that decomposes a low-rank adapter into shared and task-specific components. CoLoRA-Med freezes the components learned for previous tasks and guides new updates according to their relationship with prior ones: updates aligned with related tasks are encouraged to reuse shared directions, conflicting updates are steered into separate subspaces, and changes to parameters important for earlier tasks are regularized. We evaluate CoLoRA-Med on 11 medical visual question answering datasets across seven imaging modalities and five tasks: classification, multi-label classification, detection, cell counting, and report generation. Compared with the average of five low-rank continual learning baselines, CoLoRA-Med improves final average normalized performance by 3% and reduces mean normalized forgetting from 10.2% to 6.9%, a 32% relative reduction. These results show that combining shared adaptation with selective task specialization effectively balances plasticity and stability. Hence, CoLoRA-Med moves medical AI toward systems that keep learning new clinical skills without losing the ones clinicians already rely on. Code and checkpoints will be released.

1 Introduction

Medical visual question answering (VQA) provides a common interface for tasks that pair medical images with questions or instructions and produce textual answers (Li et al., 2023). This interface can accommodate outputs ranging from diagnostic labels and counts to locations and descriptive reports (Tu et al., 2024; Zhang et al., 2024). Although these tasks differ in visual evidence, supervision, and response structure, they often rely on overlapping clinical knowledge (Nejati Manzari et al., 2026). A general vision-language model (VLM) should therefore reuse knowledge across tasks while preserving task-specific capabilities (Tu et al., 2024), without requiring a ground-truth task identity at inference, which simplifies deployment and enables autonomous task selection.

Learning such heterogeneous tasks sequentially, however, can lead to *catastrophic forgetting*, where adaptation to new tasks degrades performance on previously learned ones (Kirkpatrick et al., 2017; Chen et al., 2024). Continual learning (CL) therefore requires balancing *plasticity*, the ability to learn new tasks, with *stability*, the ability to preserve previous capabilities (Mermillod et al., 2013; De Lange et al., 2022). This problem is particularly relevant when only the current task's data are available, preventing direct joint training and making retention a central concern (Smith et al., 2023). An empirical study of heterogeneous medical VQA shows that retention varies with learning strategy, highlighting the difficulty of accommodating different objectives within one model (Shaaban et al., 2026). These findings motivate adaptation mechanisms that account for both shared and task-specific knowledge.

Low-rank adaptation (LoRA) provides a parameter-efficient basis for this goal by training small matrix factors while freezing the backbone (Hu et al., 2022). Existing CL approaches protect se-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

lected parameters, separate successive task subspaces, or compress accumulated updates (Wang et al., 2023; Chen et al., 2025; Meng et al., 2026). While these strategies address retention, they do not fully determine how new tasks should interact with previous adaptations. Strong separation can restrict knowledge reuse, whereas unrestricted sharing can disrupt earlier capabilities. Consequently, the remaining challenge is to preserve task-specific adaptations while enabling shared learning and selective interaction.

We propose **Co**ntinual **Med**ical Learning with **Lo**w-**R**ank **A**daptation (**CoLoRA-Med**), which partitions LoRA adapters into shared and task-specific private components. During each task, only the shared and current private components are trained, while completed (i.e., previous) private components are preserved. Similarity between private updates guides conditional alignment or input-factor orthogonality, and an importance-weighted penalty constrains changes to shared parameters. At inference, an input-conditioned router selects a private component based on the question and combines it with the shared component, without requiring task identity. The main contributions are as follows:

- We introduce CoLoRA-Med, a continual adaptation framework for heterogeneous medical tasks that combines trainable shared LoRA factors with preserved task-specific factors and uses question-based routing to select the appropriate private component at inference.
- We develop a conditional interaction mechanism based on low-rank update similarity, using alignment or input-factor orthogonality together with importance-weighted regularization to balance knowledge sharing and preservation.
- We evaluate CoLoRA-Med extensively on 11 medical datasets across seven imaging modalities and five tasks: classification, multi-label classification, detection, cell counting, and report generation. Our evaluation includes comparisons with six CL baselines, retention analysis, and component ablations.

2 Related Work

Continual LoRA methods mitigate forgetting through subspace constraints, selective regularization, or structured adapter composition. O-LoRA and InfLoRA confine new updates to orthogonal or interference-free subspaces (Wang et al., 2023; Liang & Li, 2024). Following elastic weight consolidation (Kirkpatrick et al., 2017), RegLoRA regularizes entries selected by the magnitudes of earlier updates (Chen et al., 2025). IncLoRA merges a fresh per-task adapter into the backbone before learning the next task (Li et al., 2026). Other methods restructure adapter capacity: SD-LoRA decouples update magnitude and direction (Wu et al., 2025); ProgLoRA, BranchLoRA, and D-MoLE progressively add or allocate LoRA branches and experts (Yu et al., 2025; Zhang et al., 2025; Ge et al., 2025); and PCLR alternates among compression, integration, and learning over a rank pool (Meng et al., 2026).

Recent methods separate shared from task-specific capacity: CL-LoRA pairs distillation-preserved shared with task-specific adapters (He et al., 2025); HiDe-LLaVA fuses task-general components and expands task-specific ones based on layer-wise representation similarity (Guo et al., 2025); and LiLoRA shares the LoRA input factor and regularizes a shared basis using the similarity of consecutive task-specific residuals (Che et al., 2026). When task labels are unavailable at inference, task-specific parameters are selected by routing or gating: L2P matches queries to learned prompt keys (Wang et al., 2022); GainLoRA constrains gates over accumulated adapters (Liang et al., 2025); and MoE-Adapters, CL-MoE, and SMoLoRA specialize expert routing for multimodal CL (Yu et al., 2024; Huai et al., 2025; Wang et al., 2025).

Medical continual adaptation has mainly addressed domain- or modality-incremental settings with a single objective, such as multi-domain radiology report generation (Sun et al., 2024) and retinal vision-language pre-training over incrementally arriving modalities (Yao et al., 2026). CL across heterogeneous medical VQA tasks, whose outputs range from class labels to bounding boxes, counts, and free-text reports, remains underexplored. MedQwen evaluates the robustness of routed medical LoRA experts to sequential adaptation (Nejati Manzari et al., 2026), and an empirical study shows that retention depends on the CL strategy and task order (Shaaban et al., 2026). To the best of our knowledge, CoLoRA-Med is the first method to couple preserved task-private factors with importance-protected shared factors, using the similarity between private updates to decide whether

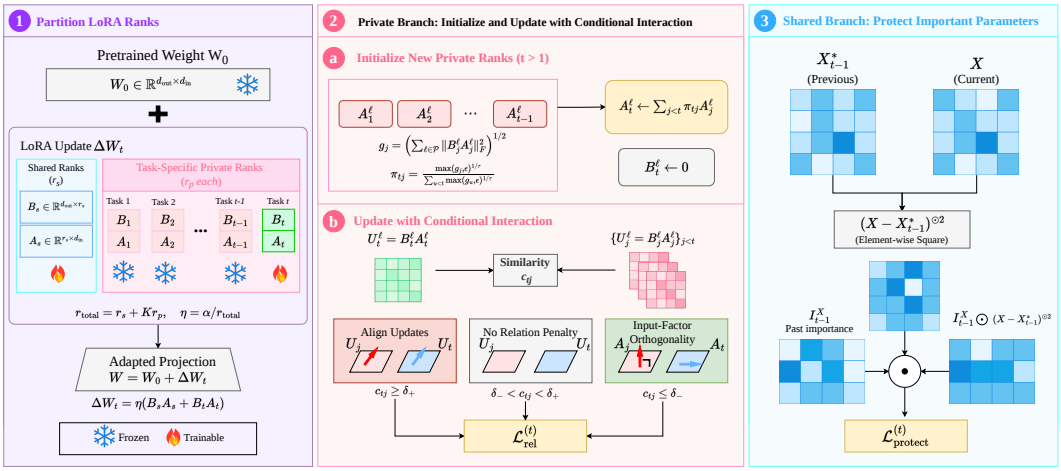


Figure 1: **Overview of CoLoRA-Med. (1) Rank partitioning:** each LoRA update is split into shared factors (A_s, B_s) and task-specific private factors (A_t, B_t). The pretrained weight and previous private factors are frozen, while shared and current private factors remain trainable. **(2) Private branch: (a)** A_t is initialized from previous private factors and B_t to zero; **(b)** similarity between the current and previous private updates determines alignment, no penalty, or input-factor orthogonality, yielding $\mathcal{L}_{\text{rel}}$. **(3) Shared branch:** important shared parameters are protected with an importance-weighted penalty, yielding $\mathcal{L}_{\text{protect}}$.

to align them, keep them orthogonal, or leave them unconstrained, thereby mitigating catastrophic forgetting across heterogeneous medical VQA tasks.

3 Preliminaries

Problem definition. We consider an ordered sequence of K tasks with datasets $\mathcal{D}_1, \ldots, \mathcal{D}_K$, all expressed in a common question–answer format. Each example $(v, q, a) \in \mathcal{D}_t$ consists of one or more medical images v, a question or instruction q, and a textual reference answer a. Given the images and the question, the model generates a text answer, which encodes a structured task output or a free-text report. At stage t, the model is trained on $\mathcal{D}_t$ and then evaluated on all tasks seen so far. The goal is to learn the current task while retaining performance on earlier ones; any degradation on earlier tasks caused by subsequent training is referred to as forgetting. In our setting, task boundaries and identities are known during training, and each stage has access only to the current task's examples. At inference, the model receives no ground-truth task identity.

Low-rank adaptation. LoRA (Hu et al., 2022) adapts a frozen weight matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ through two trainable factors, $A \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B \in \mathbb{R}^{d_{\text{out}} \times r}$:

$$W = W_0 + \Delta W, \qquad \Delta W = \frac{\alpha}{r} BA. \tag{1}$$

Here r is the adapter rank and α scales the update. The update has rank at most r and requires $r(d_{\text{in}} + d_{\text{out}})$ trainable parameters per projection, compared with $d_{\text{in}} d_{\text{out}}$ for full fine-tuning of the same weight.

4 Methodology

4.1 Shared and private rank

For each adapted projection layer, we partition the LoRA factors into shared rank r_s and private rank r_p per task:

$$\begin{aligned} r_{\text{total}} &= r_s + K r_p, \\ A &= [A_s^\top, A_1^\top, \ldots, A_K^\top]^\top, \quad B = [B_s, B_1, \ldots, B_K]. \end{aligned} \tag{2}$$

Here $A_s \in \mathbb{R}^{r_s \times d_{\text{in}}}$ and $B_s \in \mathbb{R}^{d_{\text{out}} \times r_s}$; private factors have analogous shapes with r_p in place of r_s. The effective update for task t is

$$\Delta W_t = \eta(B_s A_s + B_t A_t), \qquad \eta = \alpha / r_{\text{total}}. \tag{3}$$

Only the shared and current task's private factors receive gradients. Inactive private factors are saved, zeroed during the stage, and restored at its end. Re-zeroing them after each optimizer step prevents them from contributing to the forward computation. The active update has rank at most $r_s + r_p$, while the total allocated rank is $r_s + K r_p$ and grows linearly with the number of tasks K. See panel 1 in Figure 1.

4.2 Initialization and conditional interaction

For $t > 1$, let $\mathcal{P}$ denote the set of adapted projection layers. We measure each earlier private update's magnitude and form initialization weights:

$$g_j = \left(\sum_{\ell \in \mathcal{P}} \| B_j^\ell A_j^\ell \|_F^2 \right)^{1/2}, \qquad \pi_{tj} = \frac{\max(g_j, \epsilon)^{1/\tau}}{\sum_{u<t} \max(g_u, \epsilon)^{1/\tau}}. \tag{4}$$

We set $A_t^\ell \leftarrow \sum_{j<t} \pi_{tj} A_j^\ell$ and $B_t^\ell \leftarrow 0$. Thus initialization reuses input factors without initially copying a previous output update. For the first task, A_1 uses Kaiming-uniform initialization and $B_1 = 0$. We use $\tau = 1$ and $\epsilon = 10^{-12}$. This initialization is heuristic: earlier tasks with larger update magnitudes receive greater weight.

During training, we compare the directions of current and earlier private updates to guide regularization. We first compute their Frobenius inner product directly from the low-rank factors:

$$\langle B_t A_t, B_j A_j \rangle_F = \sum_{a=1}^{r_p} \sum_{b=1}^{r_p} (B_t^\mathsf{T} B_j)_{ab} (A_t A_j^\mathsf{T})_{ab}. \tag{5}$$

Then, these similarities are averaged over sampled projection layers to obtain a task-pair score:

$$c_{tj}^\ell = \text{clip}_{[-1,1]} \left(\frac{\langle B_t^\ell A_t^\ell, B_j^\ell A_j^\ell \rangle_F}{\max(\| B_t^\ell A_t^\ell \|_F \| B_j^\ell A_j^\ell \|_F, \epsilon)} \right), \quad c_{tj} = \frac{1}{|\mathcal{P}_{\text{sim}}|} \sum_{\ell \in \mathcal{P}_{\text{sim}}} c_{tj}^\ell, \tag{6}$$

where c_{tj}^ℓ measures directional agreement between private updates at layer ℓ, and c_{tj} averages this similarity over sampled projection layers $\mathcal{P}_{\text{sim}}$ to reduce computation. Positive and negative scores indicate aligned and opposing updates, respectively, guiding the choice of regularization. The regularizer is

$$\begin{aligned} \mathcal{L}_{\text{rel}}^{(t)} = \sum_{j<t} \Bigg[& \mathbf{1}[c_{tj} \geq \delta_+] \lambda_f (1 - c_{tj}) \\ & + \mathbf{1}[c_{tj} \leq \delta_-] \frac{\lambda_o}{r_p^2} \sum_{\ell \in \mathcal{P}} \| A_t^\ell (A_j^\ell)^\mathsf{T} \|_F^2 \Bigg]. \end{aligned} \tag{7}$$

Here $\delta_+ > 0$ and $\delta_- < 0$ are the similarity thresholds for activating alignment and input-factor orthogonality, respectively; λ_f and λ_o weight the corresponding penalties. Similarity at or above δ_+ activates alignment; similarity at or below δ_- activates input-factor orthogonality; intermediate values activate neither. See panel 2 in Figure 1.

4.3 Shared-parameter protection

To limit changes to shared parameters that were important for earlier tasks, we penalize their movement from the values saved after the previous task. See panel 3 in Figure 1. We estimate importance by accumulating squared gradients during training: $G_t^X = \sum_b (g_{t,b}^X)^{\odot 2}$ for each shared factor tensor X. Here $g_{t,b}^X$ is its gradient at backward pass b, including regularization terms, and $\odot$ denotes elementwise operations. After each task, we normalize these scores by their mean and combine them with earlier scores:

$$\widehat{I}_t^X = \frac{G_t^X}{\max(\text{mean}(G_t^X), \epsilon)}, \qquad I_1^X = \widehat{I}_1^X, \qquad I_t^X = \gamma I_{t-1}^X + (1-\gamma) \widehat{I}_t^X \quad (t > 1). \tag{8}$$

The decay factor γ controls how much past importance is retained. During task $t > 1$, we use the previous scores to weight the protection penalty:

$$\mathcal{L}_{\text{protect}}^{(t)} = \lambda_p \sum_{X \in \Theta_s} \text{mean}\left[I_{t-1}^{X} \odot (X - X_{t-1}^{*})^{\odot 2} \right]. \tag{9}$$

Here Θ_s is the set of shared factor tensors, X_{t-1}^{*} is the value saved after task $t-1$, and λ_p controls the penalty strength. Changes to entries with higher importance incur a larger penalty. The penalty is zero on the first task. The final training loss is the sum of the supervised token-level cross-entropy loss $\mathcal{L}_{\text{sup}}^{(t)}$ computed over reference answer tokens, the conditional alignment and orthogonality penalty $\mathcal{L}_{\text{rel}}^{(t)}$, and the shared-parameter protection penalty $\mathcal{L}_{\text{protect}}^{(t)}$:

$$\mathcal{L}_{\text{total}}^{(t)} = \mathcal{L}_{\text{sup}}^{(t)} + \mathcal{L}_{\text{rel}}^{(t)} + \mathcal{L}_{\text{protect}}^{(t)}. \tag{10}$$

4.4 Question-based routing

At inference, we select a task's private factors from the wording of the input question. A fixed signed hash encodes individual words, short word sequences, and boundary features into a unit-length vector $\phi(q) \in \mathbb{R}^{d_q}$, where d_q is the feature dimension. For each learned task j, we cluster these question vectors using spherical clustering and store the normalized cluster representatives $k_{j,m}$, without retaining the source questions. After learning t tasks, we select the task whose representative is most similar to the input question:

$$\widehat{j}(q) = \arg\max_{1 \le j \le t} \max_{m} \phi(q)^{\mathsf{T}} k_{j,m}. \tag{11}$$

The selected task $\widehat{j}(q)$ supplies the private factors, which are combined with the current shared factors as in Eq. (3); all other private factors are set to zero.

5 Experimental Setup

5.1 Datasets

We use public medical imaging datasets with VQA annotations released by FLARE (FLARE-MedFM, 2025). Each example pairs one or more medical images with a question or instruction and a textual reference answer. The datasets span five tasks, learned in the following order: classification, multi-label classification, detection, cell counting, and report generation. After filtering, the training data comprise 38,723 examples from 16 datasets, and the evaluation data comprise 5,561 examples from 11 datasets covering seven imaging modalities. Appendix D provides a per-dataset breakdown.

5.2 Training and baselines

We use two backbones, Qwen3-VL-4B-Instruct and InternVL3.5-4B, both with four-bit quantized weights and LoRA applied to the attention query, key, value, and output projections. As an offline reference, joint supervised fine-tuning (Joint SFT) trains on the union of all training data. We compare against six continual baselines: Sequential, O-LoRA, RegLoRA, LiLoRA, PCLR, and InCLoRA; Appendix A summarizes our implementation details.

5.3 Evaluation

After each task, we evaluate the model on every task learned so far. Model outputs are parsed according to the ground-truth task, and the corresponding task metric is then computed. We report balanced accuracy (BAcc) for classification, macro-F1 for multi-label classification, F1 averaged over intersection-over-union thresholds of 0.30, 0.50, and 0.75 for detection, mean absolute error (MAE) for cell counting, and GREEN for report generation. Each task score is the average of its per-dataset metrics, weighted by the number of valid samples.

Table 1: Final performance on public-validation examples. Scores are in % except for cell-counting MAE. Values are means ± sample standard deviations. Joint SFT (offline multi-task training) and Sequential (naive fine-tuning without forgetting mitigation) are reference points. Bold and underlined values indicate the best and second-best means among LoRA-based CL methods, ranked separately within each backbone.

Method	Class. BAcc ↑	Multi-label macro-F1 ↑	Detection F1 ↑	Cell counting MAE ↓	Report GREEN ↑
Qwen3-VL-4B-Instruct					
Joint SFT	39.18±0.17	32.38±0.07	19.74±0.10	298.43±3.14	69.02±1.26
Sequential	29.39±2.21	26.27±0.39	11.62±2.07	369.80±0.49	67.86±2.86
O-LoRA	**37.68±0.35**	29.58±0.07	15.97±0.15	401.75±0.84	**69.97±2.37**
RegLoRA	36.45±0.47	30.45±0.11	15.71±0.70	398.45±3.03	69.92±3.08
LiLoRA	33.79±2.86	25.21±0.72	11.77±1.27	351.01±23.83	69.66±2.72
PCLR	35.20±0.07	30.32±0.71	15.51±0.15	400.13±1.51	69.65±2.43
IncLoRA	35.66±0.62	30.56±0.59	16.41±0.56	400.03±2.35	69.76±2.46
CoLoRA-Med (ours)	37.06±0.97	**35.47±0.35**	**16.95±0.00**	**345.51±5.59**	68.68±2.35
InternVL3.5-4B					
O-LoRA	**36.97±0.29**	35.81±0.18	7.83±0.31	368.69±15.48	68.60±0.72
RegLoRA	36.72±0.10	36.25±0.48	13.94±0.30	344.02±6.68	74.81±0.26
LiLoRA	35.33±0.49	**38.12±1.36**	7.18±0.31	366.94±0.24	68.04±0.32
PCLR	36.65±0.32	34.15±0.07	**14.33±1.35**	348.24±2.88	74.84±0.05
IncLoRA	36.40±0.14	36.02±0.68	13.91±0.63	347.35±5.30	74.96±0.10
CoLoRA-Med (ours)	35.67±0.26	37.24±2.88	11.95±0.12	**329.89±0.06**	**75.79±0.33**

5.4 Continual learning metrics

Let $R_{t,i}$ denote the primary metric of task i after training through task $t \geq i$, and let $d_i = 1$ for higher-is-better metrics and $d_i = -1$ for cell-counting MAE. We normalize each score by the method's own performance when the task was first learned, $S_{t,i} = 1 + d_i(R_{t,i} - R_{i,i})/|R_{i,i}|$; that is, $S_{t,i} = R_{t,i}/R_{i,i}$ for higher-is-better metrics and $S_{t,i} = 2 - R_{t,i}/R_{i,i}$ for MAE. A normalized score of 1 thus indicates no change in performance. All metrics below are reported after the final task, $T = K = 5$.

Average normalized performance (ANP). $\mathrm{ANP}_t = \frac{1}{t}\sum_{i=1}^{t} S_{t,i}$ is the mean normalized score over all tasks learned so far. We report ANP_T; higher values indicate better retention. Because ANP is a direction-adjusted retention index rather than an aggregate task score, it can exceed 100%.

Worst-task retention. The worst-task retention, $\min_i S_{T,i}$, is the lowest normalized score across tasks after training. Higher values indicate better retention of the least-retained task.

Backward transfer (BWT). $\mathrm{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1}(S_{T,i} - 1)$ is the average change in earlier-task performance relative to when each task was first learned. Positive values indicate improvement, and negative values indicate degradation.

Forgetting (F). $\mathrm{F} = \frac{1}{T-1}\sum_{i=1}^{T-1}\max(0, \max_{i \leq k < T} S_{k,i} - S_{T,i})$ is the average drop from each earlier task's best normalized score before the final stage to its final score; improvements count as zero forgetting. Lower values indicate less forgetting. Both BWT and F exclude the final task, which is not followed by further training.

6 Results

6.1 Final task performance

Table 1 reports final performance on each task. On Qwen, CoLoRA-Med achieves the best mean among LoRA-based CL methods on three of the five tasks. The largest gain is in multi-label classification: its macro-F1 of 35.47% exceeds that of the next-best LoRA-based method (IncLoRA,

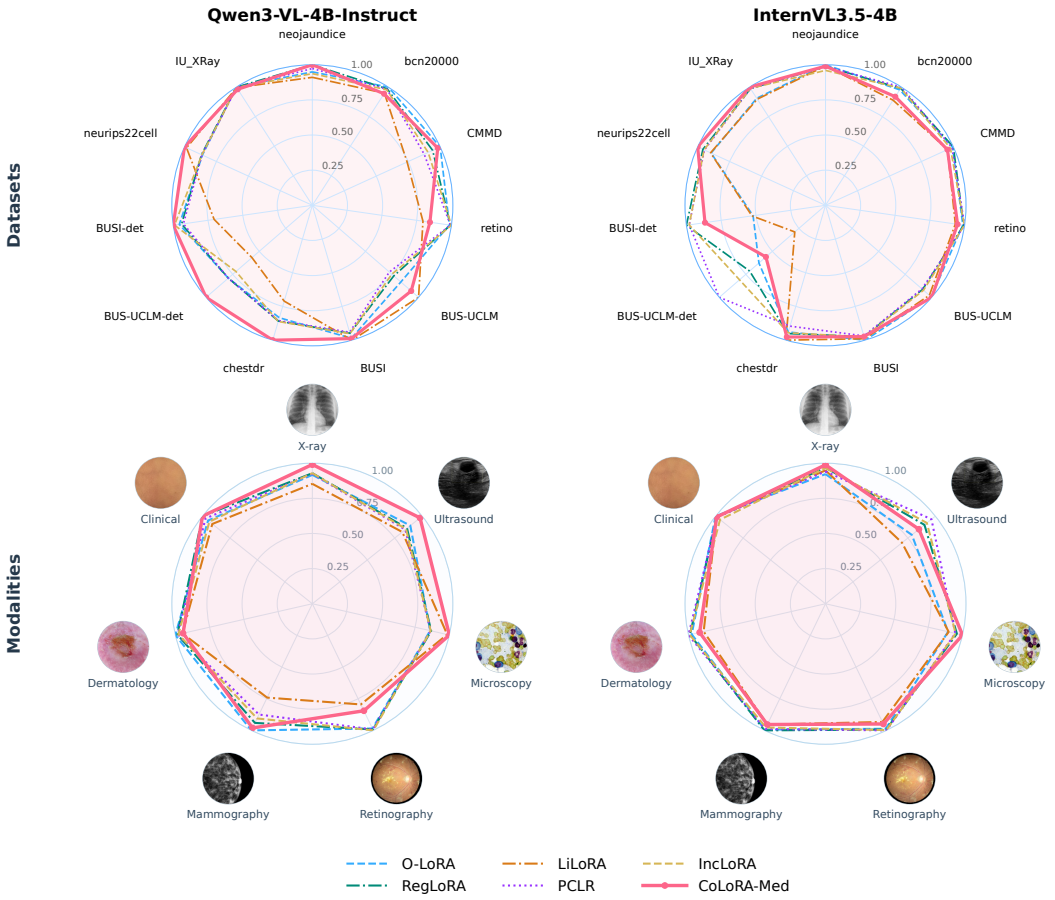


Figure 2: Final performance by dataset (top) and modality (bottom) for Qwen3-VL-4B-Instruct (left) and InternVL3.5-4B (right). Each axis is normalized by the best displayed mean within the backbone. For cell counting, we plot the best MAE divided by each method's MAE, so that larger values are better on every axis. Modality scores are unweighted means of the normalized dataset scores (four ultrasound datasets, two X-ray datasets, and one dataset for each remaining modality).

30.56%) by 4.91 percentage points (pp) and also surpasses Joint SFT (32.38%). It also leads detection (16.95% versus 16.41% F1 for IncLoRA) and cell counting (345.51 versus 351.01 MAE for LiLoRA). O-LoRA performs best on classification (37.68% balanced accuracy) and report generation (69.97% GREEN), although report scores vary little across the LoRA-based methods (68.68–69.97% GREEN). Joint SFT, which trains on all tasks simultaneously, remains a reference upper bound for all continual methods on classification, detection, and cell counting.

On InternVL, CoLoRA-Med achieves the lowest cell-counting MAE (329.89) and the highest report GREEN (75.79% versus 74.96% for IncLoRA), and ranks second in multi-label classification (37.24%). O-LoRA leads classification (36.97%), LiLoRA leads multi-label classification (38.12%), and PCLR leads detection (14.33%). For all methods, rankings depend on the task and the backbone; for example, LiLoRA obtains the highest multi-label score on InternVL but the lowest on Qwen.

6.2 Dataset and Modality Results

Figure 2 compares dataset- and modality-level profiles; full results are reported in Appendix C. On Qwen, CoLoRA-Med achieves the highest mean on five of the 11 datasets (neojaundice, chestdr, BUS-UCLM-det, BUSI-det, and neurips22cell), with the largest margin on chestdr (4.91 pp), and ranks second on CMMD, BUS-UCLM, and BUSI. At the modality level, it obtains the highest normalized score for X-ray, ultrasound, microscopy, and clinical images, the last tied with RegLoRA.

Table 2: Normalized CL metrics (%) on tasks, as defined in Section 5.4. Values are means ± sample standard deviations; bold and underlined values indicate the best and second-best means.

Method	Final ANP ↑	Worst-task retention ↑	Backward transfer ↑	Forgetting↓
Sequential	83.72±0.66	66.22±10.59	−20.35±0.82	21.29±2.16
O-LoRA	<u>97.48±0.69</u>	<u>89.74±3.01</u>	<u>−3.15±0.86</u>	<u>7.05±0.66</u>
RegLoRA	94.89±1.21	85.53±4.96	−6.39±1.52	10.57±1.57
LiLoRA	94.90±0.79	84.54±1.85	−6.38±0.98	10.61±2.39
PCLR	94.30±0.01	85.97±1.34	−7.13±0.02	12.42±1.51
IncLoRA	96.27±0.64	**90.35±1.79**	−4.67±0.80	10.35±0.81
CoLoRA-Med (ours)	**98.54±1.94**	85.89±4.69	**−1.82±2.42**	**6.94±1.69**

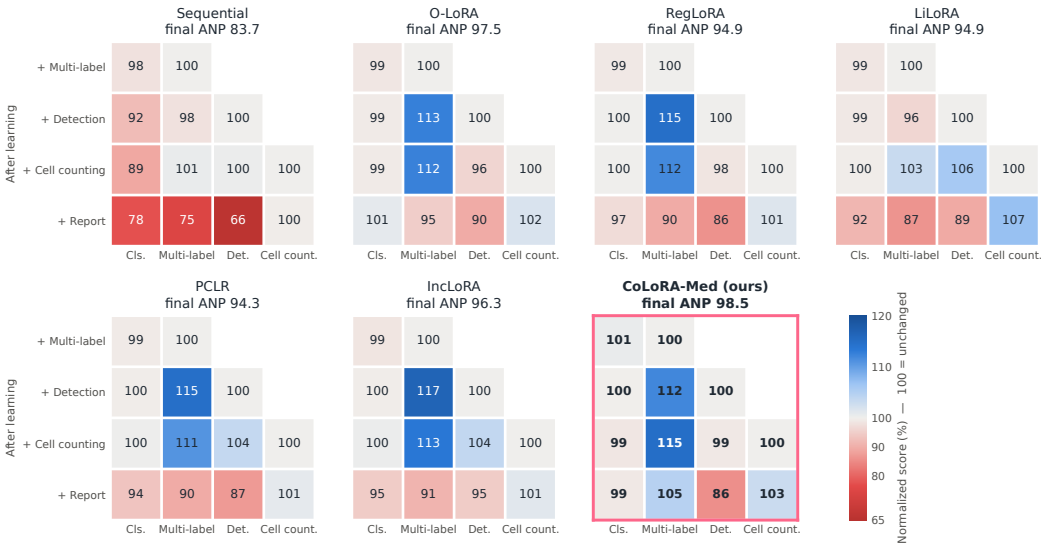


Figure 3: Stage-wise retention on Qwen3-VL-4B-Instruct. Each cell shows the normalized score $S_{t,i}$ (Section 5.4, in %) of an earlier task (column) after a later task (row) has been learned. A value of 100 indicates unchanged performance; red denotes forgetting and blue denotes improvement. Averaging the bottom row together with the final task gives the final ANP.

On InternVL, CoLoRA-Med leads the X-ray and microscopy modalities. No method is best on every dataset: for all methods, the relative ranking varies with the task and the imaging modality.

6.3 Retention

Table 2 summarizes retention on Qwen. CoLoRA-Med attains the highest final ANP (98.54%), the highest backward transfer (−1.82%), and the lowest forgetting (6.94%), followed by O-LoRA (97.48% ANP and 7.05% forgetting). Averaged over the five LoRA-based baselines (excluding Sequential), final ANP is $\bar{P}_B = 95.57\%$ and forgetting is $\bar{F}_B = 10.20\%$. CoLoRA-Med is thus 3.0 pp higher in ANP (3.1% relative, $100(P_{\text{ours}}/\bar{P}_B - 1)$) and 3.3 pp lower in forgetting (32.0% relative, $100(1 - F_{\text{ours}}/\bar{F}_B)$). Worst-task retention depends on which task each method retains least and is highest for IncLoRA and O-LoRA. Moreover, Figure 3 breaks down retention by training stage. Most forgetting occurs in the final report-generation stage, where all baselines lose 5–25 pp on multi-label classification, with Sequential dropping the most. CoLoRA-Med instead preserves classification and multi-label performance, with detection (85.9%) as its main source of forgetting.

6.4 Component ablations

We ablate CoLoRA-Med on Qwen with one epoch per task. We remove shared protection, alignment, or orthogonality by setting its loss weight to zero ("no interaction losses" removes the last

Table 3: Normalized CL metrics (%) for the component ablations. Values are means ± sample standard deviations; bold and underlined values indicate the best and second-best means.

Variant	Final ANP ↑	Worst-task retention ↑	Backward transfer ↑	Forgetting↓
Full CoLoRA-Med	98.10 ± 1.23	94.86 ± 2.80	-2.38 ± 1.54	$\mathbf{6.14 \pm 1.79}$
No shared protection	$\underline{98.58 \pm 1.18}$	94.10 ± 1.57	$\underline{-1.78 \pm 1.47}$	$\underline{6.74 \pm 1.29}$
No alignment loss	$\mathbf{98.59 \pm 0.20}$	$\mathbf{95.31 \pm 0.13}$	$\mathbf{-1.76 \pm 0.25}$	7.74 ± 0.55
No orthogonality loss	97.98 ± 1.24	93.16 ± 4.64	-2.53 ± 1.55	8.05 ± 0.24
No interaction losses	98.56 ± 1.76	$\underline{94.87 \pm 2.32}$	-1.80 ± 2.19	6.85 ± 1.60
No warm start	97.39 ± 0.86	93.56 ± 2.14	-3.27 ± 1.07	8.54 ± 1.79
Isolation only	96.61 ± 0.55	90.29 ± 2.05	-4.24 ± 0.68	9.10 ± 2.15

two). Without warm start, each task starts from freshly initialized input factors and zero output factors; the isolation-only variant keeps only private isolation and routing. As Table 3 shows, the full model has the lowest forgetting (6.14%).

7 Discussion and limitations

CoLoRA-Med performs best on cell counting on both backbones and ranks first or second on multi-label classification; on Qwen, it also leads ultrasound detection. On Qwen, this task profile is accompanied by the highest ANP and the lowest forgetting, and the ablations indicate that warm start and the combination of all components reduce forgetting relative to isolation alone. On InternVL, CoLoRA-Med achieves the best cell-counting and report scores.

No single method dominates across tasks and modalities: for every method, the balance between learning new tasks and retaining earlier ones depends on the task and the imaging modality. For example, O-LoRA leads classification and report generation on Qwen, and LiLoRA leads ultrasound classification on Qwen and multi-label classification on InternVL. The ablations show a similar balance between retention objectives, with alignment mainly reducing forgetting.

Our evaluation follows a single five-task order, and differences among the leading methods often fall within one standard deviation; evaluating multiple task orders and longer sequences would further establish the robustness of these trends. Although preserved private factors protect task-specific adaptations, updates to the shared component can still influence earlier predictions, and parameter-space similarity provides a practical but indirect proxy for task relatedness. Promising directions include relating update similarity directly to changes in earlier-task performance, protecting shared parameters more selectively, and adapting interaction thresholds and penalty weights to each task or training stage. Finally, because each private factor is low-rank, per-task storage remains small, but total private capacity and the cost of the interaction penalty grow linearly with the number of tasks; merging or compressing private factors and caching similarity computations could keep both memory and training cost bounded.

8 Conclusion

We introduced CoLoRA-Med, a continual learning method that balances learning new medical tasks with retaining previously acquired knowledge. It combines shared and task-specific low-rank components, selectively aligns or separates task updates, and protects important shared parameters. Our evaluation spans diverse medical imaging modalities, datasets, and tasks, and examines both task performance and knowledge retention. The results support combining knowledge sharing with task specialization, while the variation across tasks points to opportunities for improving this balance. CoLoRA-Med is a research method, and these results do not establish its clinical safety.

AI use statement

Generative AI was used to draft sections of the paper and to polish the writing. The authors take full responsibility for the final content, including all AI-assisted text and claims.

Ethics statement

This study evaluates continual adaptation on publicly available medical imaging datasets and does not establish clinical utility or readiness for patient care. Prediction errors and forgetting could compromise reliability, and aggregate scores do not establish equitable performance across patient populations or institutions. We do not evaluate subgroup fairness, and robustness to clinical distribution shifts. Any further use of the data, code, or checkpoints must comply with the access and licensing conditions of the source datasets, and any clinical use would require separate validation.

Reproducibility statement

Sections 4 and 5 specify the method and the scoring rules, and Appendix A lists the implementation settings. Appendices C and D provide detailed results and data summaries.

References

Chang Che, Ziqi Wang, Pengwan Yang, Cheems Wang, Hui Ma, and Zenglin Shi. LoRA in LoRA: Towards parameter-efficient architecture expansion for continual visual instruction tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(24):19978–19986, 2026. doi: 10.1609/aaai.v40i24.39082. URL https://ojs.aaai.org/index.php/AAAI/article/view/39082.

Cheng Chen, Junchen Zhu, Xu Luo, Heng Tao Shen, Jingkuan Song, and Lianli Gao. CoIN: A benchmark of continual instruction tuning for multimodel large language models. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, pp. 57817–57840, 2024. doi: 10.52202/079017-1844. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6a45500d9eda640deed90d8a62742be5-Abstract-Datasets_and_Benchmarks_Track.html.

Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ho Shing Ip, and Sam Kwong. SEFE: Superficial and essential forgetting eliminator for multimodal continual instruction tuning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7982–8001. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25n.html.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446. URL https://doi.org/10.1109/TPAMI.2021.3057446.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract.html.

FLARE-MedFM. FLARE-MLLM-2D: Dataset release. Hugging Face dataset repository, 2025. URL https://huggingface.co/datasets/FLARE-MedFM/FLARE-MLLM-2D. Accessed September 19, 2026.

Chendi Ge, Xin Wang, Zeyang Zhang, Hong Chen, Jiapei Fan, Longtao Huang, Hui Xue, and Wenwu Zhu. Dynamic mixture of curriculum LoRA experts for continual multimodal instruction tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19011–19033. PMLR, 2025. URL https://proceedings.mlr.press/v267/ge25d.html.

Haiyang Guo, Fanhu Zeng, Ziwei Xiang, Fei Zhu, Da-Han Wang, Xu-Yao Zhang, and Cheng-Lin Liu. HiDe-LLaVA: Hierarchical decoupling for continual instruction tuning of multimodal large

language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13572–13586. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.666. URL https://aclanthology.org/2025.acl-long.666/.

Jiangpeng He, Zhihao Duan, and Fengqing Zhu. CL-LoRA: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30534–30544, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/He_CL-LoRA_Continual_Low-Rank_Adaptation_for_Rehearsal-Free_Class-Incremental_Learning_CVPR_2025_paper.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. CL-MoE: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19608–19617, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Huai_CL-MoE_Enhancing_Multimodal_Large_Language_Model_with_Dual_Momentum_Mixture-of-Experts_CVPR_2025_paper.html.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://doi.org/10.1073/pnas.1611835114.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, pp. 28541–28564, 2023. doi: 10.52202/075280-1240. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html.

He Li, Yuhui Zhang, Xiaohan Wang, Kaifeng Lyu, and Serena Yeung-Levy. Fine-tuning MLLMs without forgetting is easier than you think, 2026. URL https://arxiv.org/abs/2603.14493. Preprint.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23638–23647, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liang_InfLoRA_Interference-Free_Low-Rank_Adaptation_for_Continual_Learning_CVPR_2024_paper.html.

Yan-Shuo Liang, Jia-Rui Chen, and Wu-Jun Li. Gated integration of low-rank adaptation for continual learning of large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 76577–76607, 2025. doi: 10.52202/085713-2310. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/63692d8d567db671c700df5df912204a-Abstract-Conference.html.

Weicheng Meng, Jingyang Qiao, Shaohui Liu, Zhizhong Zhang, and Yuan Xie. PCLR: Progressively compressed LoRA for multimodal continual instruction tuning. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=WdP1NVSzsz.

Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in*

Psychology, 4:504, 2013. doi: 10.3389/fpsyg.2013.00504. URL https://doi.org/10.3389/fpsyg.2013.00504.

Omid Nejati Manzari, Hojat Asgariandehkordi, Taha Koleilat, Yiming Xiao, and Hassan Rivaz. Sparse spectral LoRA: Routed experts for medical VLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35351–35362, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Nejatimanzari_Sparse_Spectral_LoRA_Routed_Experts_for_Medical_VLMs_CVPR_2026_paper.html.

Mai A Shaaban, Tausifa Jan Saleem, Alaa Mohamed, Dilnaz Utemissova, Ufaq Khan, and Mohammad Yaqub. An empirical analysis of continual learning for heterogeneous medical visual question answering. *arXiv preprint arXiv:2607.12048*, 2026.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. CODA-Prompt: COntinual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11909–11919, 2023. doi: 10.1109/CVPR52729.2023.01146. URL https://doi.org/10.1109/CVPR52729.2023.01146.

Yihua Sun, Hee Guan Khor, Yuanzheng Wang, Zhuhao Wang, Hongliang Zhao, Yu Zhang, Longfei Ma, Zhuozhao Zheng, and Hongen Liao. Continually tuning a large language model for multi-domain radiology report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15005 of *Lecture Notes in Computer Science*, pp. 177–187. Springer, 2024. doi: 10.1007/978-3-031-72086-4_17. URL https://link.springer.com/chapter/10.1007/978-3-031-72086-4_17.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Anil Palepu, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S. Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical AI. *NEJM AI*, 1(3):AIoa2300138, 2024. doi: 10.1056/AIoa2300138.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.715. URL https://aclanthology.org/2023.findings-emnlp.715.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 139–149, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_Learning_To_Prompt_for_Continual_Learning_CVPR_2022_paper.html.

Ziqi Wang, Chang Che, Qi Wang, Yangyang Li, Zenglin Shi, and Meng Wang. SMoLoRA: Exploring and defying dual catastrophic forgetting in continual visual instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 177–186, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_SMoLoRA_Exploring_and_Defying_Dual_Catastrophic_Forgetting_in_Continual_Visual_ICCV_2025_paper.html.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. SD-LoRA: Scalable decoupled low-rank adaptation for class incremental learning. In *The Thirteenth International Conference on Learning Representations (ICLR)*, pp. 58337–58353, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/92f43b1d33fae4aa1958f75317f0cec1-Abstract-Conference.html.

Yuang Yao, Ruiqi Wu, Yi Zhou, and Tao Zhou. Continual retinal vision-language pre-training upon incremental imaging modalities. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15964 of *Lecture Notes in Computer Science*, pp. 111–121. Springer, 2026. doi: 10.1007/978-3-032-04971-1_11. URL https://link.springer.com/chapter/10.1007/978-3-032-04971-1_11.

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 (CVPR)*, pp. 23219–23230, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yu_Boosting_Continual_Learning_of_Vision-Language_Models_via_Mixture-of-Experts_Adapters_CVPR_2024_paper.html.

Yahan Yu, Duzhen Zhang, Yong Ren, Xuanle Zhao, Xiuyi Chen, and Chenhui Chu. Progressive LoRA for multimodal continual instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2779–2796. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.143. URL https://aclanthology.org/2025.findings-acl.143/.

Duzhen Zhang, Yong Ren, Zhong-Zhi Li, Yahan Yu, Jiahua Dong, Chenxing Li, Zhilong Ji, and Jinfeng Bai. Enhancing multimodal continual instruction tuning with BranchLoRA. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5743–5756. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.287. URL https://aclanthology.org/2025.acl-long.287/.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024. doi: 10.1038/s41591-024-03185-2. URL https://doi.org/10.1038/s41591-024-03185-2.

A IMPLEMENTATION DETAILS

Model and optimization. We use Qwen3-VL-4B-Instruct (`Qwen/Qwen3-VL-4B-Instruct`) and InternVL3.5-4B (`OpenGVLab/InternVL3_5-4B-HF`) and report results separately for each, as means and sample standard deviations over two seeds. Following QLoRA (Dettmers et al., 2023), base weights are stored in four-bit NormalFloat with double quantization, and computation uses bfloat16. LoRA (dropout 0.05) is applied to the query, key, value, and attention output projections. We use fused AdamW with a learning rate of 10^{-5}, cosine decay, 0.03 warm-up ratio, 0.01 weight decay, gradient clipping at 1.0, and an effective batch size of 16 (batch size one with 16 accumulation steps). Each task is trained for three epochs on Qwen and, under a reduced budget, one epoch on InternVL; within each backbone, all methods share the same schedule and settings for a fair comparison. For CoLoRA-Med, we set $\delta_+ = 0.2$, $\delta_- = -0.1$, $\lambda_f = 0.05$, $\lambda_o = 0.1$, $\gamma = 0.9$, and $\lambda_p = 0.5$. The question router uses $d_q = 2048$ word unigram, bigram, trigram, and boundary features and stores up to four cluster representatives per task.

Input handling. Inputs follow the model's native chat template: the images and the original question form the user message, the reference answer forms the assistant message, and no additional system prompt is used. The loss is computed only on answer tokens. Images are converted to RGB, and the processors use a pixel range of 784–50,176 and a maximum text length of 1,024 tokens.

Baseline implementations. Sequential updates a single rank-16 adapter throughout the curriculum. We do not include a separate LoRA per task, since complete isolation would prevent tasks from sharing images and clinical knowledge that overlap across them. Our O-LoRA implementation penalizes the mean absolute overlap between input factors with a coefficient of 0.5 and the squared factor magnitude with a coefficient of 0.01. RegLoRA protects the largest 1% of cumulative-update entries, capped at 8,192 newly selected entries per layer and stage, with a coefficient of 1.0; the selected indices accumulate across tasks. Our PCLR implementation compresses the accumulated factors to rank 16 by singular-value decomposition and trains 16 new ranks within a rank-32 budget, with an input-factor orthogonality coefficient of 0.1. LiLoRA shares a rank-16 input factor and a base output factor across tasks, adds a nested rank-8 private residual with a learned fusion gate for each task, and applies the original shared-basis stability loss with a weight of 1.0. Because task identity is unavailable at inference, the private residual is selected by cosine similarity between the question embedding and stored task prototypes. IncLoRA trains a fresh rank-16 component at each stage and restores earlier factors after each optimizer step; summing these updates is equivalent to accumulating them at inference without merging them into the quantized backbone.

CoLoRA-Med uses a shared rank of $r_s = 8$ and a private rank of $r_p = 8$ per task, giving a total allocated rank of $r_{\text{total}} = 8 + 5 \times 8 = 48$ for the five-task curriculum. Only the shared factors and the selected private factors contribute to each forward pass, so the active update has rank at most $r_s + r_p = 16$, and earlier private factors remain frozen during training. The trainable rank at each stage is therefore 16, matching the rank-16 adapters of Sequential, O-LoRA, RegLoRA, and IncLoRA and the 16 new ranks trained by PCLR. Matching the total allocated rank is not possible without removing the shared/private split on which the method relies.

B LOSS-WEIGHT SENSITIVITY

To assess sensitivity to the loss weights, we use only the first two tasks. Each method resumes from its task 1 checkpoint, is trained on task 2 for one epoch, and is then evaluated for retention on 2,371 public-validation classification examples. Table 4 reports balanced accuracy, accuracy, and the BAcc drop, defined as the absolute decrease in task 1 balanced accuracy (in pp). Unlike F in Section 5.4, this drop is not normalized.

Reducing CoLoRA-Med's protection weight from 0.5 to 0.25 yields the largest retention gain: balanced accuracy rises from 35.32% to 36.00%, and the BAcc drop falls from 0.93 to 0.24 pp. All other weight changes have smaller effects. Note that this experiment measures retention of task 1, not performance on task 2.

Table 4: Sensitivity of task 1 retention to loss weights after training on task 2. Each entry shows the default setting → the alternative setting. Balanced accuracy (BAcc) and accuracy (Acc.) are in %. The BAcc drop is the absolute decrease in task 1 balanced accuracy in percentage points (pp); unlike the forgetting metric F, it is not normalized. Bold marks the better value in each pair.

Method / loss	Weight	BAcc ↑	BAcc drop (pp) ↓	Acc. ↑
O-LoRA orthogonal	$0.5 \rightarrow 0.25$	$35.18 \rightarrow \mathbf{35.33}$	$1.07 \rightarrow \mathbf{0.91}$	$57.99 \rightarrow \mathbf{58.37}$
RegLoRA regularization	$1 \rightarrow 2$	$35.31 \rightarrow \mathbf{35.37}$	$0.97 \rightarrow \mathbf{0.91}$	$\mathbf{58.41} \rightarrow 58.08$
PCLR integration	$0.1 \rightarrow 0.05$	$35.13 \rightarrow \mathbf{35.14}$	$0.98 \rightarrow 0.98$	$\mathbf{58.33} \rightarrow 58.08$
CoLoRA-Med alignment	$0.05 \rightarrow 0.025$	$\mathbf{35.55} \rightarrow 35.53$	$\mathbf{0.70} \rightarrow 0.71$	$58.41 \rightarrow 58.41$
CoLoRA-Med orthogonal	$0.1 \rightarrow 0.05$	$35.48 \rightarrow \mathbf{35.56}$	$0.76 \rightarrow \mathbf{0.68}$	$58.63 \rightarrow 58.63$
CoLoRA-Med protection	$0.5 \rightarrow 0.25$	$35.32 \rightarrow \mathbf{36.00}$	$0.93 \rightarrow \mathbf{0.24}$	$58.41 \rightarrow \mathbf{58.84}$

Table 5: Final per-dataset results (part 1). Scores are in % except for cell-counting MAE. Values are means ± sample standard deviations. Bold and underlined values indicate the best and second-best means among CL methods.

Method	neojaundice BAcc ↑	bcn20000 BAcc ↑	CMMD BAcc ↑
Qwen3-VL-4B-Instruct			
Joint SFT	53.17±1.04	27.74±0.10	33.54±0.25
Sequential	52.44±7.04	21.75±1.24	27.50±2.92
O-LoRA	50.01±1.30	**27.35±0.68**	**34.84±0.37**
RegLoRA	52.66±1.89	26.95±0.87	32.78±0.20
LiLoRA	47.97±2.69	25.94±0.00	25.83±7.01
PCLR	51.27±0.67	26.85±0.15	30.48±0.09
IncLoRA	49.36±1.30	26.74±0.54	31.56±0.58
CoLoRA-Med (ours)	**52.68±1.86**	25.81±0.66	34.18±1.79
InternVL3.5-4B			
O-LoRA	**50.81±0.58**	24.55±0.06	34.84±1.20
RegLoRA	50.22±0.25	24.88±0.24	**35.04±0.98**
LiLoRA	50.38±0.42	22.35±0.16	33.43±0.03
PCLR	50.22±2.31	**25.14±0.43**	34.67±1.33
IncLoRA	48.77±2.87	24.91±0.35	34.24±0.85
CoLoRA-Med (ours)	50.30±0.70	23.09±0.21	33.42±0.59

C Detailed results

Question-routing accuracy. After the full five-task curriculum, the router selects the correct task for at least 5,559 of the 5,561 public-validation examples (99.96%). The only errors route classification questions to cell counting.

Dataset-level performance. Tables 5, 6, 7 report dataset-level results after the complete curriculum. The classification datasets cover clinical images (neojaundice), dermatology (bcn20000), mammography (CMMD), retinography (retino), and ultrasound (BUS-UCLM and BUSI). The remaining datasets cover chest radiographs (chestdr and IU_XRay), ultrasound lesion detection (BUS-UCLM-det and BUSI-det), and microscopy cell counting (neurips22cell).

Table 6: Final per-dataset results (part 2). Scores are in % except for cell-counting MAE. Values are means ± sample standard deviations. Bold and underlined values indicate the best and second-best means among CL methods.

Method	retino BAcc ↑	BUS-UCLM BAcc ↑	BUSI BAcc ↑
Qwen3-VL-4B-Instruct			
Joint SFT	37.61±0.87	72.57±4.58	79.55±0.71
Sequential	31.20±1.33	38.65±2.50	40.66±3.93
O-LoRA	34.36±1.48	61.99±2.49	70.64±1.34
RegLoRA	34.62±1.36	57.06±1.05	67.99±0.27
LiLoRA	27.73±2.79	**74.41±4.40**	**71.65±0.45**
PCLR	34.61±0.30	53.94±1.01	67.42±0.00
IncLoRA	**34.87±1.26**	55.68±0.64	68.18±0.54
CoLoRA-Med (ours)	29.47±0.76	69.26±0.49	70.83±2.14
InternVL3.5-4B			
O-LoRA	46.64±0.45	59.76±2.21	**52.34±0.80**
RegLoRA	46.95±0.43	54.85±2.14	50.95±2.59
LiLoRA	44.10±0.27	58.68±3.63	51.77±2.68
PCLR	**47.25±0.60**	54.50±3.62	50.69±2.23
IncLoRA	47.15±0.74	55.37±2.88	51.14±2.86
CoLoRA-Med (ours)	44.86±1.08	**60.01±1.00**	50.95±2.23

Table 7: Final per-dataset results (part 3). Scores are in % except for cell-counting MAE. Values are means ± sample standard deviations. Bold and underlined values indicate the best and second-best means among CL methods.

Method	chestdr F1 ↑	BUS-UCLM-det F1 ↑	BUSI-det F1 ↑	neurips22cell MAE ↓	IU_XRay GREEN ↑
Qwen3-VL-4B-Instruct					
Joint SFT	32.38±0.07	7.54±0.56	23.96±0.05	298.43±3.14	69.02±1.26
Sequential	26.27±0.39	1.59±0.00	15.09±2.79	369.80±0.49	67.86±2.86
O-LoRA	29.58±0.07	4.42±1.70	19.97±0.38	401.75±0.84	**69.97±2.37**
RegLoRA	30.45±0.11	4.42±0.57	19.62±1.14	398.45±3.03	69.92±3.08
LiLoRA	25.21±0.72	3.17±1.12	14.74±1.31	351.01±23.83	69.66±2.72
PCLR	30.32±0.71	4.42±0.57	19.35±0.00	400.13±1.51	69.65±2.43
IncLoRA	30.56±0.59	4.02±0.00	20.70±0.76	400.03±2.35	69.76±2.46
CoLoRA-Med (ours)	**35.47±0.35**	**5.58±1.08**	**20.88±0.38**	**345.51±5.59**	68.68±2.35
InternVL3.5-4B					
Method	chestdr F1 ↑	BUS-UCLM-det F1 ↑	BUSI-det F1 ↑	neurips22cell MAE ↓	IU_XRay GREEN ↑
O-LoRA	35.81±0.18	5.42±0.59	8.66±0.22	368.69±15.48	68.60±0.72
RegLoRA	36.25±0.48	6.17±1.75	**16.63±1.00**	344.02±6.68	74.81±0.26
LiLoRA	**38.12±1.36**	2.50±0.00	8.80±0.41	366.94±0.24	68.04±0.32
PCLR	34.15±0.07	**8.64±2.91**	16.29±0.82	348.24±2.88	74.84±0.05
IncLoRA	36.02±0.68	7.00±1.75	16.30±0.24	347.35±5.30	74.96±0.10
CoLoRA-Med (ours)	37.24±2.88	4.85±0.04	14.41±0.15	**329.89±0.06**	**75.79±0.33**

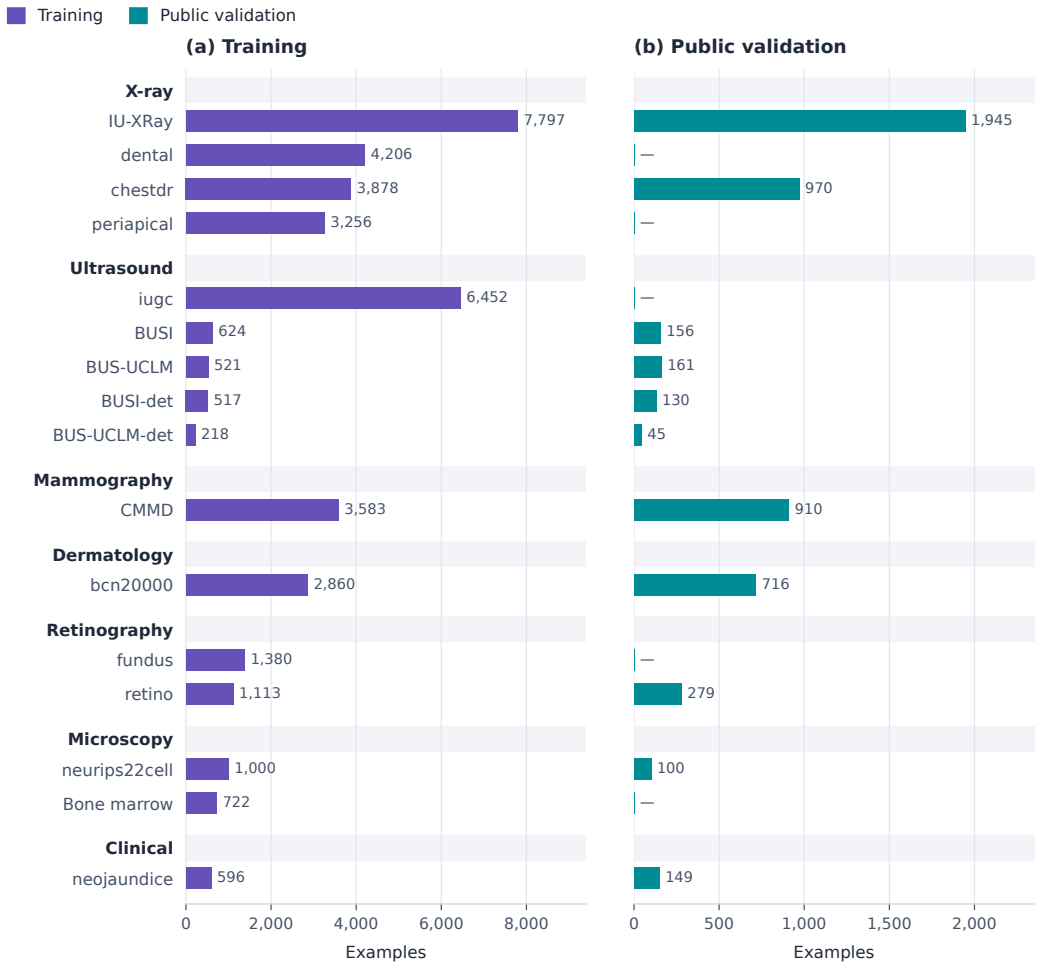


Figure 4: **Training exposure and public-validation demand.** Number of supervised examples per dataset for (a) training and (b) public validation, grouped by modality. Both axes are linear, with separate ranges; dashes mark datasets without public-validation annotations.

D TRAINING AND PUBLIC-VALIDATION DATA LANDSCAPE

Figure 4 compares dataset coverage across the two splits, Figure 5 shows their task and modality composition, and Figure 6 compares question and reference-answer lengths.

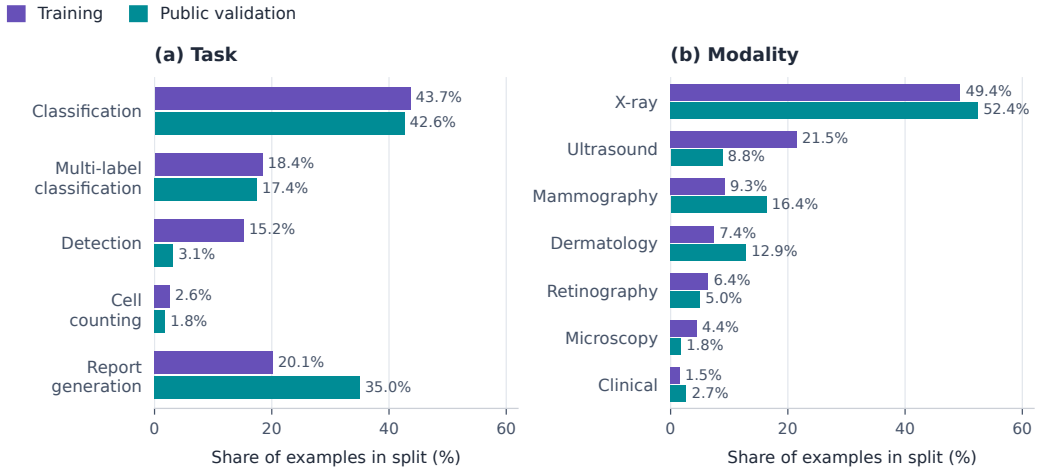


Figure 5: **Task and modality composition.** Share of supervised examples in each split by (a) task and (b) modality, out of 38,723 training and 5,561 public-validation examples.

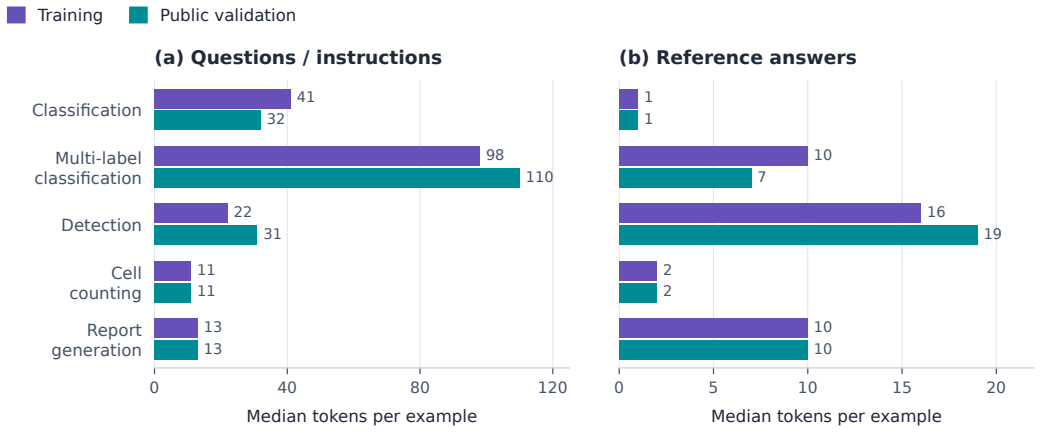


Figure 6: **Question and answer lengths by task.** Median token counts per example for (a) questions and (b) reference answers in the training and public-validation splits, shown on linear axes with separate ranges. Lengths are computed with the Qwen3-VL-4B-Instruct tokenizer, without truncation, special tokens, chat templates, or visual tokens.