FROM DISCRETE SPECIALISTS TO CONTINUOUS EXPERTISE: CONTINUAL ADAPTATION OF MEDICAL MLLMS VIA GENERATIVE LORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical multimodal large language models (MLLMs) should continually acquire capabilities across heterogeneous clinical tasks while retaining prior expertise. Existing parameter-efficient approaches face a fundamental trade-off: a single shared adapter risks interference, whereas expanding pools of task-level specialists increase storage and require expert routing. Both represent expertise at a fixed granularity, limiting scalable adaptation to individual clinical cases. We introduce **C-MAGE**, a framework that moves from discrete specialists toward continuous expertise by learning a shared generative rule for input-specific adaptation. Conditioned on image-instruction features, clinical department embeddings, and layer and projection identifiers, a shared generator predicts compact coefficients that combine shared basis matrices into instance-specific LoRA factors for the query and value projections of a frozen MLLM. This formulation integrates clinical priors, case-specific variation, and parameter reuse within a unified adaptation space. The framework is trained end-to-end through response prediction on current data and a small task-balanced replay memory, requiring neither pre-optimized expert checkpoints nor expert routing at inference. For evaluation, we construct **MedCLAS**, a clinically organized benchmark comprising 151,187 samples from 13 datasets across six departments and unifying classification, visual question answering, and report generation through a common instruction-following interface. Experiments demonstrate strong performance across heterogeneous medical tasks, showing that continual adaptation can proceed by refining shared generative expertise rather than accumulating discrete specialists. Code will be made publicly available.

1 INTRODUCTION

Medical multimodal large language models (MLLMs) offer a unified interface for interpreting medical images and following clinical instructions, supporting disease classification, visual question answering, and report generation. However, they are typically trained on fixed datasets, while clinical data arrive progressively across specialties, modalities, and institutions. Continual adaptation requires acquiring new capabilities while retaining existing ones. Beyond distribution shift, successive tasks may demand different visual evidence, clinical concepts, and response formats within the same instruction-following interface. Effective continual adaptation thus requires both knowledge sharing across heterogeneous tasks and specialization for individual clinical inputs.

Conventional continual learning mitigates forgetting via replay (Castro et al., 2018; Hou et al., 2019), regularization (Kirkpatrick et al., 2017), and architectural adaptation (Abati et al., 2020; Yoon et al., 2017). For MLLMs, substantial forgetting under sequential instruction tuning (Chen et al., 2024; Guo et al., 2025b) has motivated parameter-efficient approaches built on LoRA (Hu et al., 2022), including separable routing in SMoLoRA (Wang et al., 2025), expert selection in MR-LoRA (Zhao et al., 2025), and hierarchical adaptation in HiDe-LLaVA (Guo et al., 2025a). These approaches highlight a tension in organizing expertise: shared adaptation parameters enable reuse but remain susceptible to interference (Fig. 1(a)), whereas expanding specialist pools preserves isolated parameter states at the cost of growing storage and difficult expert routing (Fig. 1(b)). Moreover, task boundaries only coarsely capture adaptation needs: cases within a task may require different specializations, while

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.

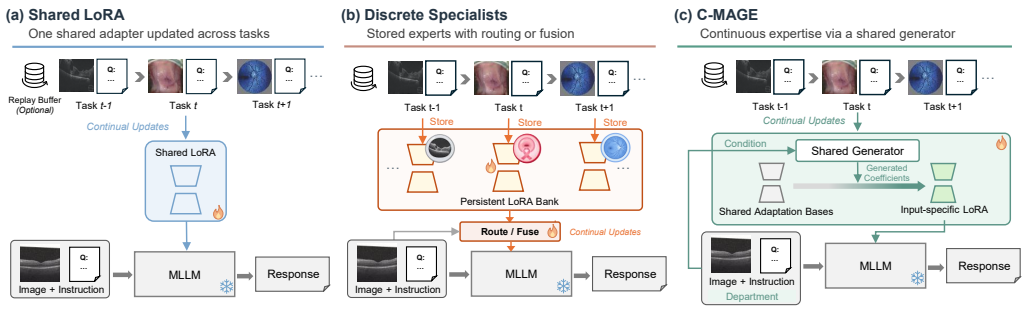


Figure 1: Conceptual comparison of continual adaptation paradigms for medical MLLMs. (a) A shared LoRA is updated across tasks. (b) Discrete LoRA specialists are stored and routed or fused for each input. (c) C-MAGE continually learns a shared generator that produces input-specific LoRA factors conditioned on the image, instruction, and department context.

cases across tasks may benefit from shared adaptations. This raises a fundamental question: *must continually accumulated expertise be represented as a collection of discrete static specialists?*

Moving beyond discrete specialists requires a framework that represents expertise continuously rather than as fixed task-level modules (Fig. 1(c)). Such a framework should generate adaptations at the input level, reuse knowledge across related inputs, and learn directly from sequential task supervision without relying on pre-collected experts or expert routing. Yet unconstrained parameter generation may produce unstable adaptations and offer little structure for knowledge sharing as the generator evolves. The central challenge is therefore to organize input-specific updates within a reusable adaptation space that supports both clinical specialization and continual knowledge retention.

To make this challenge concrete and evaluate it systematically, we construct **MedCLAS**, a benchmark for Medical Continual Learning Across Specialties (Fig. 2). MedCLAS organizes 13 datasets into a continual-learning stream spanning six clinical service units—radiology, pathology, dermatology, ophthalmology, gynecology, and gastroenterology—and ten imaging modalities. It contains 151,187 samples and 98,621 images, and unifies classification, visual question answering, and report generation through a common image-instruction-response interface. Each dataset defines one continual-learning task, while department membership provides coarse clinical context shared by related tasks. Models are evaluated on all observed tasks after each update, enabling systematic assessment of capability acquisition and retention across heterogeneous clinical domains, modalities, dataset scales, and response requirements. To our knowledge, MedCLAS provides one of the longest and most heterogeneous publicly available continual instruction-tuning streams for medical MLLMs.

Motivated by this setting, we introduce **C-MAGE** (Fig. 3), a framework for continual medical adaptation of MLLMs through generative expertise. The framework leverages the natural hierarchy of clinical practice: department metadata provides a coarse clinical prior, while the image and instruction specify what expertise is needed for the current case. Accordingly, C-MAGE combines multimodal input features with a learnable department embedding to condition a shared hypernetwork (Ha et al., 2017). The hypernetwork predicts layer- and projection-specific coefficient matrices, which are expanded through shared basis matrices into instance-specific LoRA factors for the query and value projections of a frozen MLLM. The shared bases define reusable adaptation subspaces shared across inputs and transformer layers, while the generated coefficients express case-specific variation within these subspaces. This factorization reduces the dimensionality of parameter generation and enables diverse adaptations within a common basis-constrained family, without retaining separate task-specific LoRA experts. Continual learning jointly refines the generator, conditioning embeddings, and shared bases through autoregressive response prediction on current data and a small dataset-balanced replay memory. Task supervision thereby shapes the generated adaptations end-to-end. At inference, C-MAGE generates LoRA factors from the image-instruction input and department context, holds them fixed throughout response decoding, and discards them afterward. Experiments on MedCLAS demonstrate strong performance against conventional and recent continual-learning methods, together with robustness across task orders, replay budgets, LoRA ranks, and vision-language backbones. These results highlight the potential of shifting continual medical adaptation from accumulating discrete specialists to progressively refining a shared generative rule.

Our contributions are threefold:

- We introduce C-MAGE, a framework that explores the transition from discrete task specialists to a continuous family of input-specific adaptations for continual medical MLLMs.
- We develop a structured generative LoRA parameterization that uses observed department context to guide specialization and shared adaptation bases to support parameter reuse, generating instance-specific updates within a common adaptation space.
- We construct MedCLAS, a service-unit-organized benchmark for continual medical instruction tuning across six clinical units and three task types. Experiments demonstrate promising performance compared to state-of-the-art continual-learning approaches.

2 Related Work

Continual Learning. Continual learning (CL) aims to acquire knowledge from sequential data while retaining previously learned capabilities. Replay-based methods revisit past examples (Castro et al., 2018; Hou et al., 2019), regularization-based methods protect learned behavior or important parameters, as exemplified by LwF (Li & Hoiem, 2017) and EWC (Kirkpatrick et al., 2017), and architecture-based methods isolate or expand task-specific components (Abati et al., 2020; Yoon et al., 2017). For pretrained models, parameter-efficient fine-tuning (PEFT) provides lightweight adaptation mechanisms through prompts (Lester et al., 2021), adapters (Gao et al., 2024), and LoRA (Hu et al., 2022). These mechanisms support continual learning through selective updating, reuse, or expansion of adaptation parameters. Extending them to MLLMs additionally requires preserving visual understanding and instruction-following capabilities as multimodal tasks evolve.

Continual Learning for MLLMs. Recent benchmarks reveal substantial forgetting during sequential multimodal instruction tuning (Chen et al., 2024; Guo et al., 2025b), motivating methods for protecting and organizing lightweight adaptations. ModalPrompt uses dual-modality guidance to select and fuse task-specific prompts (Zeng et al., 2025). Within LoRA-based approaches, RegLoRA regularizes important entries in previously learned updates (Chen et al., 2025), while KeepLoRA constrains adaptation to subspaces orthogonal to pretrained and previous-task knowledge directions (Luo et al., 2026). Other methods emphasize expert organization: MoELoRA combines low-rank experts through input-dependent gating (Chen et al., 2024), SMoLoRA separates visual- and instruction-oriented routing (Wang et al., 2025), and MR-LoRA uses an MLLM-based router to select isolated experts (Zhao et al., 2025). ProgLoRA progressively expands an adapter pool with task-aware allocation and recall (Yu et al., 2025), whereas HiDe-LLaVA combines task-specific upper-layer adaptation with task-general lower-layer fusion (Guo et al., 2025a). These methods primarily protect learned parameter updates or organize stored adaptation modules; expanding expert pools additionally introduces storage growth and expert-selection requirements. C-MAGE explores a complementary direction in how expertise is represented: a shared generator synthesizes instance-specific LoRA factors within reusable adaptation subspaces, conditioned on multimodal inputs and hospital department context. Continual learning thus refines a common generative parameterization, supported by a small replay memory, without retaining a separate LoRA expert for each task.

3 MedCLAS Benchmark

We construct the **MedCLAS** benchmark to evaluate continual adaptation across heterogeneous medical tasks. As summarized in Fig. 2, MedCLAS integrates 13 datasets from six clinical departments, covering ten imaging modalities and three task types: classification, visual question answering (VQA), and report generation. It contains 151,187 samples and 98,621 images, with 118,525 training samples and 32,662 test samples. Each dataset defines one continual-learning task, while department membership provides coarse clinical context shared by related tasks.

The radiology unit comprises COVID-CXP (Wang et al., 2023; Kermany et al., 2018), SV-CT&X-Ray (Liu et al., 2021; Lau et al., 2018), IU-X-Ray (Pavlopoulos et al., 2019), and SVR-MRI (Liu et al., 2021; Lau et al., 2018; Butsanets et al., 2025). Pathology includes PCam (Veeling et al., 2018) and PathVQA (He et al., 2020); dermatology includes HAM-Skin8, constructed from public skin-lesion datasets (Tschandl et al., 2018; Codella et al., 2018; Hernández-Pérez et al., 2024), and

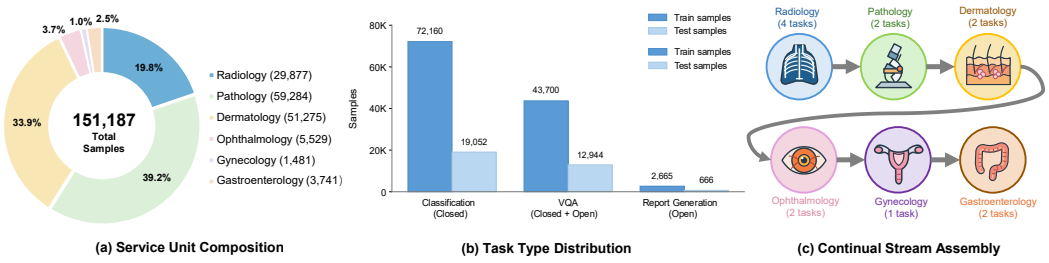


Figure 2: MedCLAS benchmark composition and continual-learning stream. (a) Sample distribution across six clinical departments. (b) Training and test sample counts by task type. (c) Default department-ordered assembly of the 13-task continual-learning stream.

Fitzpatrick (Groh et al., 2021). Ophthalmology comprises Yangxi (Liu et al., 2019) and OCT-C8 (Subramanian et al., 2022), while gynecology includes Cervical (Payette et al., 2017). Gastroenterology comprises Kvasir-VQA-Sub (Gautam et al., 2024) and HyperKvasir-Sub (Borgli et al., 2020). Dataset composition and preprocessing are detailed in Appendix A.

All tasks use a unified image-instruction-response interface: classification generates textual category labels, VQA generates answers to image-conditioned questions, and report generation produces clinical descriptions. Human-assisted verification checks metadata, labels, prompts, and split consistency, and identifies potential image overlap between the training and test sets. Tasks are organized by clinical department, and the default stream begins with radiology because it spans the broadest range of modalities and all three task types, while the remaining departments follow a fixed but otherwise arbitrary order without implying clinical priority. Models are evaluated on all observed tasks after each update, enabling assessment of capability acquisition and retention across a clinically structured stream with diverse modalities, response formats, and dataset sizes.

4 Methodology

C-MAGE reformulates continual adaptation as learning a shared rule for generating expertise, rather than accumulating discrete LoRA specialists. As illustrated in Fig. 3, C-MAGE combines image-instruction features with a learned department embedding to capture both instance-specific requirements and coarse clinical context. A shared hypernetwork further incorporates layer and projection identifiers and predicts compact coefficient matrices, which are expanded through reusable bases into instance-specific LoRA factors for the query and value projections of the frozen MLLM. This structured parameterization enables case-adaptive specialization within a shared adaptation space. Throughout the continual stream, only the generator, conditioning embeddings, and shared bases are updated, eliminating the need to store or route among task-specific adapters.

4.1 Preliminaries

Continual learning setup. Let $\mathcal{D}_{1:T} = \{\mathcal{D}_1, \ldots, \mathcal{D}_T\}$ denote a sequence of T medical vision-language datasets, where $\mathcal{D}_t$ arrives at continual step t. Each sample is represented as $(x_i^v, x_i^q, y_i, d_i) \in \mathcal{D}_t$, where x_i^v, x_i^q, and y_i denote the medical image, instruction, and target response, respectively, and $d_i \in \{1, \ldots, N_D\}$ denotes its clinical department. Department metadata is available during training and inference as coarse clinical context rather than a task identifier. At step t, C-MAGE learns from $\mathcal{D}_t$ and a small dataset-balanced replay memory $\mathcal{M}_{t-1}$. After each update, it must generate input-specific adaptations for all observed datasets $\mathcal{D}_{1:t}$, enabling the frozen MLLM to produce responses without dataset identity, task identity, or task-specific adapters.

Frozen MLLM backbone. We adopt a LLaVA-style MLLM (Li et al., 2023) with frozen parameters $\Theta = \{\theta_{\text{vis}}, \theta_{\text{proj}}, \theta_{\text{llm}}\}$, consisting of a visual encoder, a multimodal projector, and a decoder-only language model. Given a medical image x_i^v, the visual encoder extracts $\mathbf{e}_i^v = f_{\text{vis}}(x_i^v; \theta_{\text{vis}}) \in \mathbb{R}^{L_{\text{vis}} \times d_{\text{vis}}}$, where L_{vis} is the number of visual tokens and d_{vis} is the visual hidden dimension. The projector then maps these features into the language-model embedding space: $\widetilde{\mathbf{e}}_i^v = f_{\text{proj}}(\mathbf{e}_i^v; \theta_{\text{proj}}) \in \mathbb{R}^{L_{\text{vis}} \times d_{\text{llm}}}$.

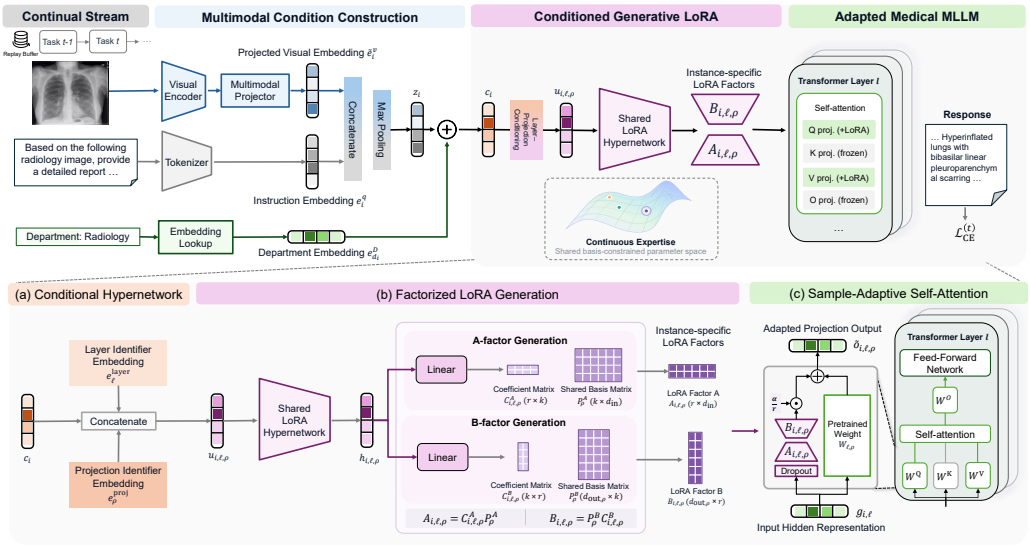


Figure 3: Overview of C-MAGE. A shared LoRA generator produces instance-specific LoRA factors from multimodal input features, learned department embeddings, and layer/projection conditions, providing sample-adaptive query and value projections while the MLLM backbone remains frozen.

Let $\mathbf{e}_i^q \in \mathbb{R}^{L_{\text{text}} \times d_{\text{llm}}}$ denote the instruction-token embeddings. The corresponding multimodal embedding sequence is $\mathbf{e}_{\text{multi}}(x_i) = [\widetilde{\mathbf{e}}_i^v; \mathbf{e}_i^q] \in \mathbb{R}^{(L_{\text{vis}}+L_{\text{text}}) \times d_{\text{llm}}}$. All backbone components remain frozen throughout continual adaptation. The multimodal embedding sequence is subsequently max-pooled into an instance-level representation, which is combined with a learnable clinical-department embedding to construct the input-specific condition for the parameter generator.

Low-rank adaptation. For a frozen linear projection $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and input representation $\mathbf{h} \in \mathbb{R}^{d_{\text{in}}}$, LoRA (Hu et al., 2022) parameterizes the adaptation as

$$\mathbf{h}' = (\mathbf{W} + \Delta\mathbf{W})\mathbf{h} = \mathbf{W}\mathbf{h} + \frac{\alpha}{r}\mathbf{B}\mathbf{A}\mathbf{h}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{r \times d_{\text{in}}}$ and $\mathbf{B} \in \mathbb{R}^{d_{\text{out}} \times r}$ are the LoRA factor matrices, r is the LoRA rank, and α is the scaling factor. Conventional LoRA directly optimizes and stores fixed factor matrices for each adapted projection, whereas C-MAGE uses a shared LoRA generator to produce input-, layer-, and projection-conditioned coefficient matrices that are expanded through shared trainable bases into instance-specific LoRA factors for the query and value projections.

4.2 Department-aware Instance-Conditioned Generative LoRA

Maintaining an ever-expanding pool of static LoRA specialists incurs growing storage costs and makes expert routing increasingly challenging. C-MAGE instead learns a shared generative rule that synthesizes instance-specific LoRA adaptations on demand. For each image-instruction pair, it combines fine-grained multimodal features with a learnable department embedding, while layer and projection embeddings identify the target adaptation site. This joint conditioning enables diverse low-rank adaptations within a shared parameterization.

Instance- and department-conditioned context. For each image-instruction pair, we obtain a fixed-dimensional instance representation by max-pooling the multimodal embedding sequence: $\mathbf{z}_i = \text{MaxPool}(\mathbf{e}_{\text{multi}}(x_i^v, x_i^q)) \in \mathbb{R}^{d_{\text{llm}}}$. While $\mathbf{z}_i$ captures input-specific information, we additionally introduce a learnable clinical-department embedding to provide coarse clinical context. Specifically, we define a trainable department embedding matrix $\mathbf{E}^D \in \mathbb{R}^{N_D \times d_{\text{llm}}}$, whose d_i-th row, $\mathbf{e}_{d_i}^D = \mathbf{E}^D[d_i]$, represents the associated clinical department d_i. The resulting conditioning representation is

$$\mathbf{c}_i = \text{sg}(\mathbf{z}_i) + \mathbf{e}_{d_i}^D, \tag{2}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator, treating the frozen multimodal representation as a fixed conditioning signal. The department embedding matrix is zero-initialized and learned jointly with the shared LoRA generator. Thus, $\mathbf{c}_i$ integrates fine-grained multimodal information with coarse clinical context for instance-specific LoRA generation.

Shared LoRA hypernetwork. To enable a single shared generator to produce distinct adaptations for different transformer layers and self-attention projection types, we associate each target layer and projection with a learnable identifier embedding. Specifically, $\mathbf{e}_\ell^{\mathrm{layer}}$ is a trainable lookup embedding indexed by layer ℓ, while $\mathbf{e}_\rho^{\mathrm{proj}}$ is a trainable lookup embedding indexed by projection type $\rho \in \{\mathrm{Q}, \mathrm{V}\}$. For sample i, the hypernetwork input is

$$\mathbf{u}_{i,\ell,\rho} = [\mathbf{c}_i; \mathbf{e}_\ell^{\mathrm{layer}}; \mathbf{e}_\rho^{\mathrm{proj}}], \tag{3}$$

where $[\cdot;\cdot]$ denotes concatenation. A shared two-layer MLP $H(\cdot)$ then transforms the representation:

$$\mathbf{h}_{i,\ell,\rho} = H(\mathbf{u}_{i,\ell,\rho}) = \mathrm{GELU}\Big(\mathbf{W}_2\,\mathrm{GELU}\big(\mathbf{W}_1\,\mathrm{LN}(\mathbf{u}_{i,\ell,\rho}) + \mathbf{b}_1\big) + \mathbf{b}_2\Big), \tag{4}$$

where $\mathbf{W}_1, \mathbf{W}_2, \mathbf{b}_1, \mathbf{b}_2$, together with the layer and projection identifier embeddings, are trainable generator parameters. Here, $\mathbf{c}_i$ provides the input-specific conditioning information, while the layer and projection embeddings specify the target adaptation site. The resulting $\mathbf{h}_{i,\ell,\rho}$ therefore conditions subsequent coefficient generation jointly on the input, transformer layer, and projection type.

Factorized LoRA generation. Directly generating full LoRA factors requires the hypernetwork to predict $r(d_{\mathrm{in}} + d_{\mathrm{out},\rho})$ values for each sample, layer, and projection, where $d_{\mathrm{out},\rho}$ is the output dimension of projection ρ. To reduce this dimensionality, two trainable linear heads map $\mathbf{h}_{i,\ell,\rho}$ to compact coefficient matrices:

$$\mathbf{C}^A_{i,\ell,\rho} = \mathrm{reshape}_{r\times k}\left(\mathbf{W}_A\mathbf{h}_{i,\ell,\rho} + \mathbf{b}_A\right), \tag{5}$$

$$\mathbf{C}^B_{i,\ell,\rho} = \mathrm{reshape}_{k\times r}\left(\mathbf{W}_B\mathbf{h}_{i,\ell,\rho} + \mathbf{b}_B\right), \tag{6}$$

where k denotes the basis dimension, and $\mathbf{W}_A, \mathbf{b}_A, \mathbf{W}_B, \mathbf{b}_B$ are trainable parameters of the coefficient prediction heads. Importantly, $\mathbf{C}^A_{i,\ell,\rho}$ and $\mathbf{C}^B_{i,\ell,\rho}$ are generated dynamically for each sample, layer, and projection rather than stored as trainable parameters. The coefficient matrices are expanded through trainable projection-specific shared basis matrices to obtain the full LoRA factors:

$$\mathbf{A}_{i,\ell,\rho} = \mathbf{C}^A_{i,\ell,\rho}\mathbf{P}^A_\rho, \qquad \mathbf{P}^A_\rho \in \mathbb{R}^{k\times d_{\mathrm{in}}}, \tag{7}$$

$$\mathbf{B}_{i,\ell,\rho} = \mathbf{P}^B_\rho\mathbf{C}^B_{i,\ell,\rho}, \qquad \mathbf{P}^B_\rho \in \mathbb{R}^{d_{\mathrm{out},\rho}\times k}. \tag{8}$$

The basis matrices $\{\mathbf{P}^A_\rho, \mathbf{P}^B_\rho\}$ are shared across samples and transformer layers but learned separately for each projection type, whereas the coefficient matrices and resulting LoRA factors vary with the input, target layer, and projection. Since $\mathbf{C}^A_{i,\ell,\rho} \in \mathbb{R}^{r\times k}$ and $\mathbf{C}^B_{i,\ell,\rho} \in \mathbb{R}^{k\times r}$, the generator predicts only $2rk$ input-dependent values, rather than the $r(d_{\mathrm{in}} + d_{\mathrm{out},\rho})$ values required for direct factor generation, while still producing full-sized LoRA factors. The shared bases define reusable adaptation subspaces, and the generated coefficients specify input-, layer-, and projection-dependent coordinates within them. By varying these coefficients, C-MAGE produces a continuous family of low-rank adaptations within a shared basis-constrained parameterization.

Sample-adaptive self-attention. Given the generated LoRA factors, the frozen query and value projections within each target self-attention layer receive sample-specific low-rank updates. Let $\mathbf{g}_{i,\ell}$ denote the hidden representation entering the corresponding projection at layer ℓ. For $\rho \in \{\mathrm{Q}, \mathrm{V}\}$,

$$\widetilde{\mathbf{o}}_{i,\ell,\rho} = \mathbf{W}_{\ell,\rho}\mathbf{g}_{i,\ell} + \frac{\alpha}{r}\mathbf{B}_{i,\ell,\rho}\mathbf{A}_{i,\ell,\rho}\,\mathrm{Dropout}(\mathbf{g}_{i,\ell}), \tag{9}$$

where the base projection $\mathbf{W}_{\ell,\rho}$ remains frozen, and the key and output projections remain unadapted. The conditioning representation and corresponding LoRA factors are computed once per image-instruction pair and held fixed throughout autoregressive decoding. The generated coefficients and factors are temporary and discarded after inference; only the shared generator, conditioning embeddings, and bases are retained across the continual stream, with no instance-, dataset-, or task-specific adapters stored. We use asymmetric initialization: the A-coefficient head is initialized with small nonzero parameters, whereas the B-coefficient head is initialized to zero. Thus, $\mathbf{B}_{i,\ell,\rho}\mathbf{A}_{i,\ell,\rho} = \mathbf{0}$ at initialization, so Eq. 9 initially recovers the frozen projection while preserving a valid gradient path to the B-coefficient head.

Training objective. At continual step t, the shared trainable parameters Φ are optimized on the current dataset and a small dataset-balanced replay memory $\mathcal{M}_{t-1}$ containing samples from previously observed datasets, forming $\mathcal{S}_t = \mathcal{D}_t \cup \mathcal{M}_{t-1}$. For a target response $y_i = (y_{i,1}, \ldots, y_{i,L_i^y})$, we minimize the autoregressive cross-entropy loss

$$\mathcal{L}_{\mathrm{CE}}^{(t)} = -\mathbb{E}_{(x_i^v, x_i^q, y_i, d_i)\sim\mathcal{S}_t}\left[\sum_{s=1}^{L_i^y} \log p_{\Theta,\Phi}\left(y_{i,s} \mid x_i^v, x_i^q, d_i, y_{i,<s}\right)\right], \tag{10}$$

where L_i^y denotes the number of tokens in the i-th response, s indexes the response-token position, and $y_{i,<s} = (y_{i,1}, \ldots, y_{i,s-1})$ denotes the preceding response tokens. The backbone parameters Θ remain frozen, while only Φ is optimized. The complete trainable parameter set is

$$\Phi = \left\{\mathbf{E}^D, \mathbf{E}^{\mathrm{layer}}, \mathbf{E}^{\mathrm{proj}}, \mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2, \mathbf{W}_A, \mathbf{b}_A, \mathbf{W}_B, \mathbf{b}_B, \{\mathbf{P}_\rho^A, \mathbf{P}_\rho^B\}_{\rho\in\{\mathrm{Q},\mathrm{V}\}}\right\}. \tag{11}$$

Here, $\mathbf{E}^{\mathrm{layer}}$ and $\mathbf{E}^{\mathrm{proj}}$ are trainable lookup matrices for the layer and projection identifiers, respectively. The coefficient matrices $\mathbf{C}^A, \mathbf{C}^B$ and resulting LoRA factors $\mathbf{A}, \mathbf{B}$ are generated on demand and are neither independent trainable parameters nor retained across inputs. Notably, C-MAGE is trained end-to-end from response supervision, without requiring pre-optimized specialist adapters or parameter-reconstruction targets, thereby avoiding the separate training and storage of specialist checkpoints. The continual training procedure is provided in Algorithm 1 in Appendix B.

4.3 Inference

At inference, C-MAGE receives an image-instruction pair and its department metadata d_i. Such coarse metadata is typically available in hospital workflows, where each examination is associated with a clinical service unit, and does not require dataset or task identity. C-MAGE forms the conditioning representation $\mathbf{c}_i$ using Eq. 2 and generates layer- and projection-specific coefficient matrices, which are expanded through the shared bases into instance-specific LoRA factors for the query and value projections. These factors are generated once from the initial input, held fixed throughout autoregressive decoding, and discarded afterward. C-MAGE therefore enables on-demand adaptation without storing or routing among task-specific LoRA experts.

5 Experiments

Baseline. We compared C-MAGE with sequential LoRA finetuning (Hu et al., 2022), replay (Hou et al., 2019), regularization methods (Li & Hoiem, 2017; Kirkpatrick et al., 2017), and recent MLLM continual-learning methods based on prompt routing (Zeng et al., 2025), mixture-of-expert adaptation (Chen et al., 2024; Wang et al., 2025; Zhao et al., 2025), LoRA subspace constraints (Chen et al., 2025; Luo et al., 2026), and progressive expert expansion (Yu et al., 2025). 'Finetune LoRA Each', which selects independently trained experts using ground-truth task identity, serves as a task-ID oracle reference. Detailed descriptions of all comparison methods are provided in Appendix C.1.

Implementation. We implemented C-MAGE on LLaVA-Med-v1.5 (Li et al., 2023) and trained it on one NVIDIA A100 GPU with 80GB memory. Training used AdamW with no weight decay, a cosine learning-rate schedule, a warm-up ratio of 0.03, and BF16 mixed precision. Unless otherwise specified, the learning rate was 2×10^{-4}. The pretrained MLLM remained frozen, with generated LoRA updates applied to the query and value projections using rank $r = 16$ and scaling factor $\alpha = 16$. The hypernetwork had a hidden width of 512, layer- and projection-identifier embedding dimensions of 48 and 24, respectively, and a shared basis dimension of $k = 128$. Dataset-specific training epochs are provided in Appendix C.2, Table 3.

Evaluation metrics. We evaluated classification using accuracy, open-ended VQA using recall, closed-ended VQA using exact-match accuracy, and IU-X-Ray report generation using F1 and BLEU-4. We report the final **Avg. Score** over the 17 non-BLEU metric slots and **BWT** to quantify knowledge retention, where negative BWT indicates forgetting. Full definitions are provided in Appendix C.3.

5.1 Comparison with State-of-the-Art Methods

Table 1 compares C-MAGE with state-of-the-art methods on MedCLAS under the primary experimental order. Existing methods did not consistently preserve the task competence acquired throughout

Table 1: Quantitative comparison on MedCLAS under the task order Rad → Path → Derm → Oph → Gyn → Gas. 'Finetune LoRA Each' uses oracle task identity to select independently trained experts. Avg. Score averages all non-BLEU task metrics, and BWT is computed over the same metric set.

Method	Radiology							Pathology		
	COVID-CXP	SV-CT&X-Ray		IU-X-Ray		SVR-MRI		PCam	PathVQA	
	ACC	Open	Closed	F1	BLEU	Open	Closed	ACC	Open	Closed
Finetune LoRA Each	88.11	82.13	79.94	41.44	4.90	73.18	84.97	86.59	37.44	91.54
LoRA-FT (Hu et al., 2022)	26.84	18.25	44.59	1.73	0.00	16.28	40.97	2.04	0.54	42.06
Replay (Hou et al., 2019)	70.23	57.43	70.27	36.05	4.16	60.72	75.08	82.47	12.34	72.90
LwF (Li & Hoiem, 2017)	28.93	4.37	47.30	0.21	0.00	3.81	49.35	41.94	0.54	50.47
EWC (Kirkpatrick et al., 2017)	27.84	4.37	40.54	0.46	0.00	5.30	43.93	0.01	1.43	28.04
ModalPrompt (Zeng et al., 2025)	46.96	12.03	47.30	0.16	0.00	14.74	50.88	60.02	0.27	57.94
MoELoRA (Chen et al., 2024)	51.20	39.10	54.05	0.46	0.00	38.92	59.76	51.50	3.88	63.55
RegLoRA (Chen et al., 2025)	43.81	10.98	35.14	0.01	0.00	9.42	33.67	49.85	0.54	22.43
SMoLoRA (Wang et al., 2025)	61.78	49.21	60.47	13.57	0.60	41.19	53.80	66.74	11.38	66.03
MRLoRA (Zhao et al., 2025)	56.74	38.62	68.92	32.00	4.46	45.56	68.38	48.90	12.63	65.42
KeepLoRA (Luo et al., 2026)	41.25	37.04	51.35	0.47	0.00	35.63	61.93	64.60	4.16	49.53
ProgLoRA (Yu et al., 2025)	84.72	19.25	15.41	1.43	0.00	27.20	32.75	50.04	2.28	20.23
C-MAGE (Ours)	77.06	60.46	71.80	37.98	4.51	65.16	78.21	83.91	16.62	78.74

Method	Dermatology			Ophthalmology		Gynecology	Gastroenterology		Analysis	
	HAM-Skin8	Fitzpatrick		Yangxi	OCT-C8	Cervical	Kvasir-VQA	HyperKvasir	Overall Results	
	ACC	Open	Closed	ACC	ACC	ACC	ACC	ACC	Avg. Score↑	BWT↑
Finetune LoRA Each	77.35	54.70	91.07	100.00	91.64	60.81	89.71	70.54	76.54	-
LoRA-FT (Hu et al., 2022)	2.36	5.13	62.30	59.67	25.38	16.76	29.77	50.14	26.17	-36.10
Replay (Hou et al., 2019)	63.98	28.01	76.23	100.00	89.56	54.66	82.00	69.89	64.81	1.17
LwF (Li & Hoiem, 2017)	4.89	0.01	43.44	64.33	51.35	50.50	34.91	61.61	31.64	-31.98
EWC (Kirkpatrick et al., 2017)	13.87	0.02	40.16	70.50	28.42	16.55	26.50	47.32	23.25	-38.53
ModalPrompt (Zeng et al., 2025)	32.47	8.74	51.64	99.17	39.30	56.01	41.03	69.32	40.47	-16.93
MoELoRA (Chen et al., 2024)	37.52	8.10	77.87	85.33	84.00	35.00	70.14	66.07	48.61	-15.74
RegLoRA (Chen et al., 2025)	10.49	1.28	2.46	72.17	53.41	50.50	50.22	46.43	28.99	-35.38
SMoLoRA (Wang et al., 2025)	28.68	21.92	76.18	100.00	89.36	48.99	71.05	69.57	54.70	-8.07
MRLoRA (Zhao et al., 2025)	54.18	7.31	54.10	98.83	69.59	50.50	62.80	51.02	52.09	-3.46
KeepLoRA (Luo et al., 2026)	16.90	8.46	51.64	58.90	8.28	16.89	22.30	53.57	34.29	-26.77
ProgLoRA (Yu et al., 2025)	10.75	2.88	2.00	74.33	8.46	17.57	7.62	50.89	25.17	-15.53
C-MAGE (Ours)	67.72	33.63	84.60	100.00	81.76	60.47	85.53	75.00	68.16	-6.34

MedCLAS. As observed, sequential LoRA finetuning only achieved an Avg. Score of 26.17 with a BWT of -36.10, showing severe interference when heterogeneous updates were accumulated within a single adaptation state (Table 1). LwF and EWC similarly produced strongly negative BWT values, indicating that constraining shared updates alone is insufficient over a long and heterogeneous medical stream. Recent MLLM-specific methods improved performance on particular datasets or reduced forgetting, but remained substantially below the task-ID oracle, suggesting that prompt selection, LoRA regularization, or routing among discrete experts addresses only part of the task-agnostic continual-learning problem. Replay provides an especially informative comparison. It achieved a positive BWT of 1.17 but a final Avg. Score of 64.81, showing that limited forgetting does not necessarily imply strong continual learning when the competence acquired at individual updates is itself modest. Notably, C-MAGE achieved a final Avg. Score of 68.2 after the complete stream. Instead of maintaining an expanding pool of LoRA experts, it learns a shared generative rule that synthesizes input-specific LoRA factors from shared bases and conditioned coefficients. Generated once per input and reused throughout decoding, these adaptations eliminate persistent expert storage and inference-time routing. Thus, C-MAGE mitigates forgetting while retaining strong acquired competence, rather than benefiting from weak intermediate task learning. These results support shared generative adaptation as a scalable alternative to expanding expert banks for long-horizon continual learning in medical MLLMs.

5.2 Ablation Study

We ablate the LoRA factor generator in C-MAGE, as shown in Fig. 4(a), with implementation details and further analysis provided in Appendix Section D.1. Directly generating full instance-specific LoRA factors yields a final average score of only 34.75. Introducing shared trainable bases substantially improves the score to 65.87, while adding learned department embeddings further increases it to 68.16. These results validate the two key design choices of C-MAGE: shared-basis factorization provides a structured adaptation space for instance-specific updates, while department-aware conditioning supplies complementary clinical context.

5.3 Sensitivity Analysis

Task Order. We further evaluate C-MAGE under two shuffled task orders; the exact training sequences and dataset-level results are provided in Appendix Section D.2 and Fig. 4(b). C-MAGE

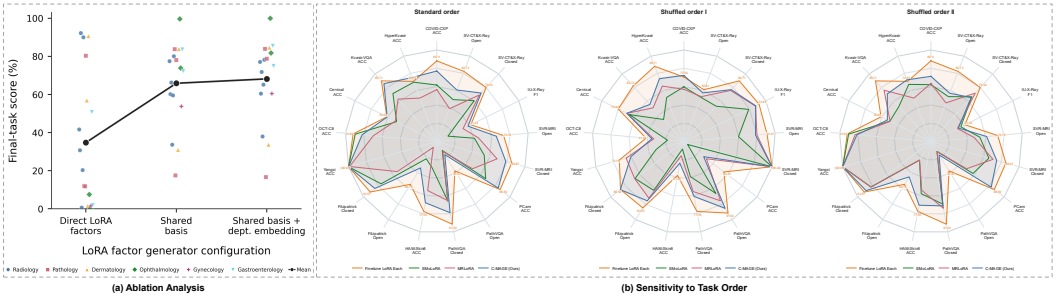


Figure 4: (a) Ablation study comparing direct factor generation, shared bases, and department-conditioned shared bases. (b) Dataset-level performance of Finetune LoRA Each (oracle), SMoLoRA, MRLoRA, and C-MAGE under the default and two shuffled task orders.

achieves final average scores of 68.16, 66.19, and 65.52 under the default and two shuffled orders, respectively. Despite moderate decreases of 1.97-2.64 points, performance remains broadly consistent across departments and datasets. These results indicate reasonable robustness to task order, while suggesting that early task exposure still shapes the generator's subsequent adaptation. Warm-starting from a general-purpose generator pretrained across tasks may provide a stronger initialization and further reduce order sensitivity, which we leave for future work.

LoRA rank, replay size, and backbone. We evaluate C-MAGE across different LoRA ranks, replay-memory sizes, and vision-language backbones, with detailed results provided in Appendix Section D.3 and Fig. 5(a)-(c). C-MAGE achieves final average scores of 65.56, 68.16, and 68.12 with LoRA ranks of 8, 16, and 32, respectively, indicating limited sensitivity once sufficient adaptation capacity is reached. We therefore adopt $r = 16$, which performs best without unnecessary rank expansion. C-MAGE also remains robust under the extreme replay constraint of $M = 1$, while the non-monotonic trend across memory sizes suggests that larger replay budgets do not necessarily improve retention. We use $M = 200$ in the main experiments for controlled comparison with baseline methods. Replacing LLaVA-Med with LLaVA (Liu et al., 2023) yields only a modest decrease from 68.16 to 67.12, suggesting that C-MAGE is not strongly tied to a specific vision-language backbone. Overall, C-MAGE remains stable across the evaluated configurations.

5.4 Functional Geometry Analysis and Training Paradigms

We investigate whether C-MAGE must reproduce independently optimized expert LoRAs. Fig. 6(a) visualizes their functional residual responses across six datasets. Although expert and C-MAGE-generated updates occupy clearly separated regions, both retain distinct dataset-dependent structure. C-MAGE therefore learns structured specialization rather than collapsing to a generic update, suggesting that effective adaptations need not follow the geometry of an expert bank. Fig. 6(b) further shows that explicitly aligning generated updates with expert transformations does not improve downstream performance. Because individual experts are optimized independently, their solutions may unnecessarily constrain a shared generator intended to transfer knowledge across tasks. Direct optimization with task loss instead allows C-MAGE to discover a distinct and more effective adaptation space through its effects on response generation. Detailed objectives and analyses are provided in Appendix D.4.

6 Conclusion

In this paper, we presented C-MAGE, a framework that explores continual medical adaptation through a shared generative rule for input-specific LoRA updates. By combining department-aware conditioning with shared adaptation bases, C-MAGE supports case-specific specialization within a reusable parameter space, without maintaining a growing bank of experts. It learns directly from response prediction with a small replay memory, avoiding reliance on pre-optimized specialist checkpoints. We also introduced MedCLAS as a heterogeneous testbed for continual medical instruction tuning. Promising empirical results support a shift from discrete specialists toward continuous expertise: accumulated capabilities can be expressed through a continually refined adaptation rule that generates specialized parameters on demand.

AI USE STATEMENT

The authors utilized LLMs as a writing assistant during the preparation of this manuscript. Their role was primarily focused on tasks related to language refinement and improving the clarity of the narrative. Their application included: (1) enhancing the conciseness and academic tone of sentences and paragraphs; (2) restructuring complex sentences for better readability; and (3) brainstorming alternative phrasings for key scientific arguments. It is important to emphasize that all core scientific ideas, experimental design, results, and conclusions were exclusively developed by the human authors. The LLM served as a sophisticated editing and brainstorming tool. All text generated or modified by the LLM was critically reviewed, edited, and ultimately approved by the authors to ensure it accurately reflected our original intent and scientific findings.

ETHICS STATEMENT

Our study does not involve human subjects, personally identifiable information, or sensitive clinical data. All datasets used are publicly available and have been properly anonymized by their providers. We have carefully considered potential risks and believe that our work poses no foreseeable harm to patients or institutions. This research strictly adheres to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

All implementation details, including model architectures, training hyperparameters, and data preprocessing steps, are described in the main paper and the Appendix. The source code, along with documentation and scripts to reproduce our experiments, will be released upon paper acceptance.

REFERENCES

Davide Abati, Jakub Tomczak, Tijmen Blankevoort, Simone Calderara, Rita Cucchiara, and Babak Ehteshami Bejnordi. Conditional channel gated networks for task-aware continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3931–3940, 2020.

Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020.

Léo Butsanets, Charles Corbière, Julien Khlaut, Pierre Manceron, and Corentin Dancette. Radimagenet-vqa: A large-scale ct and mri dataset for radiologic visual question answering. *arXiv preprint arXiv:2512.17396*, 2025.

Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 233–248, 2018.

Cheng Chen, Junchen Zhu, Xu Luo, Heng T Shen, Jingkuan Song, and Lianli Gao. Coin: A benchmark of continual instruction tuning for multimodel large language models. *Advances in neural information processing systems*, 37:57817–57840, 2024.

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. *arXiv preprint arXiv:2505.02486*, 2025.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International journal of computer vision*, 132(2):581–595, 2024.

Sushant Gautam, Andrea M Storås, Cise Midoglu, Steven A Hicks, Vajira Thambawita, Pål Halvorsen, and Michael A Riegler. Kvasir-vqa: A text-image pair gi tract dataset. In *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications*, pp. 3–12, 2024.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1820–1828, 2021.

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, 2025a.

Haiyang Guo, Fei Zhu, Hongbo Zhao, Fanhu Zeng, Wenzhuo Liu, Shijie Ma, Da-Han Wang, and Xu-Yao Zhang. Mcitlib: Multimodal continual instruction tuning library and benchmark. *arXiv preprint arXiv:2508.07307*, 2025b.

David Ha, Andrew Dai, and Quoc VL. Hypernetworks. In *International Conference on Learning Representations*, 2017.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, Noel CF Codella, Veronica Rotemberg, Allan C Halpern, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Brian Helba, et al. Bcn20000: Dermoscopic lesions in the wild. *Scientific data*, 11(1):641, 2024.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 831–839, 2019.

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*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, and J. Dong. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 3045–3059, 2021.

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. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Chi Liu, Xiaotong Han, Zhixi Li, Jason Ha, Guankai Peng, Wei Meng, and Mingguang He. A self-adaptive deep learning method for automated eye laterality detection based on color fundus photography. *Plos one*, 14(9):e0222025, 2019.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Tong Wei, and Min-Ling Zhang. Keeplora: Continual learning with residual gradient adaptation. *arXiv preprint arXiv:2601.19659*, 2026.

John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. A survey on biomedical image captioning. In *Proceedings of the second workshop on shortcomings in vision and language*, pp. 26–36, 2019.

Jack Payette, Jake Rachleff, and C De Graaf. Intel and mobileodt cervical cancer screening kaggle competition: cervix type classification using deep learning and image classification. *Stanford University*, 2017.

Malliga Subramanian, Kogilavani Shanmugavadivel, Obuli Sai Naren, K Premkumar, and K Rankish. Classification of retinal oct images using deep learning. In *2022 international conference on computer communication and informatics (ICCCI)*, pp. 1–7. IEEE, 2022.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 210–218. Springer, 2018.

Boyuan Wang, Du Zhang, and Zonggui Tian. Corotrans-cl: A novel transformer-based continual deep learning model for image recognition of coronavirus infections. *Electronics*, 12(4):866, 2023.

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*, pp. 177–186, 2025.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.

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, 2025.

Fanhu Zeng, Fei Zhu, Haiyang Guo, Xu-Yao Zhang, and Cheng-Lin Liu. Modalprompt: Towards efficient multimodal continual instruction tuning with dual-modality guided prompt. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12137–12152, 2025.

Hongbo Zhao, Fei Zhu, Haiyang Guo, Meng Wang, Rundong Wang, Gaofeng Meng, and Zhaoxiang Zhang. Mllm-cl: Continual learning for multimodal large language models. *arXiv preprint arXiv:2506.05453*, 2025.

Appendix

A MedCLAS Dataset Composition and Construction 14

B Continual Training Algorithm for C-MAGE 15

C Additional Implementation Details 15
- C.1 Comparison Methods 15
- C.2 Dataset-specific Epoch Schedules 16
- C.3 Evaluation Metrics 16

D Additional Experiments 16
- D.1 Ablation Analyses 16
- D.2 Task-Order Robustness 17
- D.3 Additional Sensitivity Experiments 17
- D.4 Functional Geometry and Training Paradigms 18

E Limitations and Societal Impacts 20

A MEDCLAS DATASET COMPOSITION AND CONSTRUCTION

As shown in Table 2, **MedCLAS** comprises 13 tasks organized into six clinical departments: radiology, pathology, dermatology, ophthalmology, gynecology, and gastroenterology. Across these departments, it covers ten imaging modalities: CT, X-ray, MRI, histopathology, macroscopy, dermoscopy, fundus photography, OCT, colposcopy, and gastrointestinal endoscopy. The tasks encompass classification, visual question answering (VQA), and report generation. Fig. 2 summarizes the departmental organization and dataset statistics. Some task collections combine or reformulate public datasets, as described below.

Dataset composition. Classification tasks span all six departments. COVID-CXP combines COVID (Wang et al., 2023) and CXP (Kermany et al., 2018) for respiratory-image classification using CT and chest X-ray images. PCam (Veeling et al., 2018) provides histopathology patches labeled for the presence of metastatic tissue, while HAM-Skin8 combines public skin-lesion datasets into a unified classification task (Tschandl et al., 2018; Codella et al., 2018; Hernández-Pérez et al., 2024). Yangxi (Liu et al., 2019) and OCT-C8 (Subramanian et al., 2022) support ophthalmic classification using fundus photography and retinal OCT, respectively. Cervical (Payette et al., 2017) provides colposcopy images, and HyperKvasir-Sub is drawn from HyperKvasir (Borgli et al., 2020) for gastrointestinal endoscopy classification. The VQA tasks cover radiology, pathology, dermatology, and gastroenterology. SV-CT&X-Ray combines CT and X-ray question-answer samples from Slake (Liu et al., 2021) and VQA-Rad (Lau et al., 2018). SVR-MRI combines MRI samples from these two sources with RadImageNet (Butsanets et al., 2025) into an MRI-focused VQA task. PathVQA (He et al., 2020) provides pathology image-question-answer samples, Fitzpatrick (Groh et al., 2021) is used to construct a dermatology VQA task, and Kvasir-VQA-Sub is drawn from Kvasir-VQA (Gautam et al., 2024) for endoscopic VQA. Report generation is represented by IU-X-Ray (Pavlopoulos et al., 2019), which pairs chest X-ray images with diagnostic reports.

Table 2: Composition of MedCLAS. The sample counts refer to image-instruction-response instances; a single image may be associated with multiple VQA instances.

Dataset	Service unit	Task	Modality	Train samples	Train images	Test samples	Test images	Answer type
COVID-CXP (Wang et al., 2023; Kermany et al., 2018)	Radiology	Classification	CT, X-ray	13,094	9,155	2,768	1,696	Closed
SV-CT&X-Ray (Liu et al., 2021; Lau et al., 2018)	Radiology	VQA	CT, X-ray	3,904	341	870	85	Closed & Open
IU-X-Ray (Pavlopoulos et al., 2019)	Radiology	Report generation	Chest X-ray	2,665	2,665	666	666	Open
SVR-MRI (Liu et al., 2021; Lau et al., 2018; Butsanets et al., 2025)	Radiology	VQA	MRI	4,777	535	1,133	130	Closed & Open
PCam (Veeling et al., 2018)	Pathology	Classification	Histopathology	26,213	26,213	6,555	6,555	Closed
PathVQA (He et al., 2020)	Pathology	VQA	Histopathology, Macroscopy	19,755	2,599	6,761	858	Closed & Open
HAM-Skin8 (Tschandl et al., 2018; Codella et al., 2018; Hernández-Pérez et al., 2024)	Dermatology	Classification	Dermoscopy	28,141	28,141	7,039	7,039	Closed
Fitzpatrick (Groh et al., 2021)	Dermatology	VQA	Dermoscopy	12,876	3,228	3,219	2,158	Closed & Open
Yangxi (Liu et al., 2019)	Ophthalmology	Classification	Fundus photography	1,205	1,156	308	290	Closed
OCT-C8-Sub (Subramanian et al., 2022)	Ophthalmology	Classification	OCT	2,042	1,640	1,974	1,584	Closed
Cervical Cancer (Payette et al., 2017)	Gynecology	Classification	Colposcopy	1,185	1,185	296	296	Closed
Kvasir-VQA-Sub (Gautam et al., 2024)	Gastroenterology	VQA	GI endoscopy	2,388	150	961	60	Open
HyperKvasir-Sub (Borgli et al., 2020)	Gastroenterology	Classification	GI endoscopy	280	140	112	56	Closed

Instruction-format harmonization. All samples are represented as image-instruction-response triples with associated department labels. Classification instructions request a category expressed as a textual response, retaining closed-set evaluation within a generative interface. VQA samples use standardized question and answer formats, preserving existing pairs where available. Report-generation instructions request clinical descriptions, with the associated reports serving as targets. This shared interface supports a common response-prediction objective across task types.

Quality verification. Human-assisted verification checks dataset metadata, modality assignments, split consistency, and label mappings. Prompts and question-answer pairs are reviewed for consistency and clinical plausibility. Image-duplication and overlap checks are included to reduce potential train-test leakage, particularly for task collections assembled from multiple sources.

Continual-learning organization. Each dataset defines one task, and tasks are grouped by clinical department to form the continual stream. This organization captures coarse clinical relationships while preserving differences in modalities, labels, and response requirements. Models are evaluated on all observed tasks after each update, enabling analysis of new capability acquisition, prior capability retention, and knowledge transfer within and across departments.

B CONTINUAL TRAINING ALGORITHM FOR C-MAGE

Algorithm 1 summarizes the continual training procedure of C-MAGE. At each task, the pretrained MLLM remains frozen, while the adaptation parameters Φ are optimized using current-task data and replay samples from previously observed datasets. For each input, pooled multimodal features and department context condition the generation of LoRA factors for the target query and value projections, while the shared generator is trained end-to-end through the response prediction objective. After training on dataset $\mathcal{D}_t$, up to M samples are retained and added to the replay memory.

Algorithm 1 Continual Training of C-MAGE

Require: Data stream $\mathcal{D}_{1:T}$; pretrained MLLM F_Θ; per-dataset replay quota M
Ensure: Learned adaptation parameters Φ

Freeze Θ; initialize trainable parameters Φ (Eq. 11)
Initialize replay memory $\mathcal{M}_0 \leftarrow \varnothing$
for $t = 1, \ldots, T$ **do**
  for each training minibatch $\mathcal{B}$ sampled from $\mathcal{D}_t \cup \mathcal{M}_{t-1}$ **do**
    Extract frozen multimodal input embeddings for all samples in $\mathcal{B}$
    Compute $\mathbf{z}_i$ by masked max pooling, excluding padding and target-response tokens
    Construct $\mathbf{c}_i$ using Eq. 2
    for each target layer ℓ and projection $\rho \in \{\mathrm{Q}, \mathrm{V}\}$ **do**
      Compute $\mathbf{h}_{i,\ell,\rho} = H([\mathbf{c}_i; \mathbf{e}_\ell^{\mathrm{layer}}; \mathbf{e}_\rho^{\mathrm{proj}}])$
      Predict $\mathbf{C}^A_{i,\ell,\rho}$ and $\mathbf{C}^B_{i,\ell,\rho}$ using Eqs. 5–6
      Generate $\mathbf{A}_{i,\ell,\rho} = \mathbf{C}^A_{i,\ell,\rho}\mathbf{P}^A_\rho$ and $\mathbf{B}_{i,\ell,\rho} = \mathbf{P}^B_\rho \mathbf{C}^B_{i,\ell,\rho}$
    end for
    Apply the generated LoRA updates using Eq. 9
    Compute $\mathcal{L}^{(t)}_{\mathrm{CE}}$ using Eq. 10 and update Φ
  end for
  Sample a replay subset $\mathcal{R}_t \subseteq \mathcal{D}_t$ with $|\mathcal{R}_t| = \min(M, |\mathcal{D}_t|)$
  Update $\mathcal{M}_t \leftarrow \mathcal{M}_{t-1} \cup \mathcal{R}_t$ ▷ $|\mathcal{M}_t| \le tM$
end for

C ADDITIONAL IMPLEMENTATION DETAILS

C.1 COMPARISON METHODS

We compare C-MAGE with conventional continual-learning strategies and recent MLLM-specific approaches. **Finetune LoRA Each** independently trains one LoRA expert for each dataset and selects the corresponding expert using ground-truth task identity at inference, serving as a task-ID oracle reference. **LoRA-FT** (Hu et al., 2022) sequentially updates a shared LoRA adapter throughout the continual stream. **Replay** (Hou et al., 2019) trains the shared adapter using current samples together with a small memory retained from previous datasets. **LwF** (Li & Hoiem, 2017) mitigates forgetting through knowledge distillation from the preceding model, whereas **EWC** (Kirkpatrick et al., 2017) regularizes changes to parameters estimated to be important for previously learned knowledge.

Among MLLM-specific approaches, **ModalPrompt** (Zeng et al., 2025) learns task-specific prompts with multimodal prompt selection and fusion. **MoELoRA** (Chen et al., 2024) employs multiple low-rank experts with input-dependent gating. **RegLoRA** (Chen et al., 2025) regularizes important components of LoRA updates to reduce interference across tasks. **SMoLoRA** (Wang et al., 2025) uses separate visual- and instruction-oriented routing paths and combines their routed LoRA outputs. **MRLoRA** (Zhao et al., 2025) maintains task-specific LoRA experts and employs an MLLM-based generative router for expert selection. **KeepLoRA** (Luo et al., 2026) constrains new LoRA updates within a residual subspace, while **ProgLoRA** (Yu et al., 2025) progressively expands the LoRA pool with task-aware allocation and retrieval. Unless otherwise stated, all methods use the same backbone, data order, and task-agnostic final evaluation protocol. Replay-based methods use the same memory budget of 200 samples per previously observed dataset.

C.2 Dataset-specific Epoch Schedules

Table 3 lists the training epochs used for each dataset.

Table 3: **Dataset-specific epoch schedules.** C-MAGE uses dataset-specific training epochs to account for differences in optimization convergence across heterogeneous medical tasks.

Radiology				Pathology		Dermatology		Ophthalmology		Gynecology	Gastroenterology	
COVID -CXP	SV-CT &X-Ray	IU -X-Ray	SVR -MRI	PCam	PathVQA	HAM -Skin8	Fitz patrick	Yangxi	OCT-C8	Cervical	Kvasir -VQA	Hyper Kvasir
1	3	20	3	1	3	1	3	3	3	3	3	6

C.3 Evaluation Metrics

Classification datasets were evaluated using accuracy. For VQA datasets, open-ended responses were evaluated using recall, while closed-ended responses were evaluated using exact-match accuracy after answer normalization. IU-X-Ray report generation was evaluated using F1 and BLEU-4, abbreviated as BLEU. BLEU is reported separately and excluded from the aggregate score because its scale is not directly comparable with the percentage-based metrics used for the other tasks.

The final average score aggregates the $K = 17$ non-BLEU metric slots with equal weight. Let $S_m^{(t)}$ denote the score of metric slot m after completing update t, and let $t(m)$ denote the update at which the dataset associated with metric m was learned. After the final update T, the average score is computed as

$$\text{Avg. Score} = \frac{1}{K} \sum_{m=1}^{K} S_m^{(T)}. \tag{12}$$

Backward transfer (BWT) measures the change in performance on previously learned datasets after subsequent continual updates. Defining $\mathcal{M}_{\text{past}} = \{m \in \{1, \ldots, K\} : t(m) < T\}$, we compute

$$\text{BWT} = \frac{1}{|\mathcal{M}_{\text{past}}|} \sum_{m \in \mathcal{M}_{\text{past}}} \left(S_m^{(T)} - S_m^{(t(m))} \right). \tag{13}$$

Metrics associated with the final update are excluded because no subsequent learning occurs. Negative BWT indicates net forgetting, whereas positive BWT indicates beneficial backward transfer.

D Additional Experiments

D.1 Ablation Analyses

Fig. 4(a) evaluates three configurations of the LoRA factor generator. In the first configuration, the generator is conditioned only on the image-instruction representation,

$$c_i = \text{sg}(z_i),$$

and directly predicts the complete instance-specific LoRA factors for every adapted layer and projection. This variant achieves an average final-task score of 34.75. Directly synthesizing high-dimensional LoRA factors places a substantial burden on a single shared generator, which must simultaneously represent reusable adaptation knowledge and capture input-specific variation. The second configuration introduces shared trainable basis matrices while retaining the same image-instruction condition. Instead of predicting complete LoRA factors, the generator predicts low-dimensional coefficients that combine the shared bases into instance-specific adaptations. This factorization increases the average score from 34.75 to 65.87, a gain of 31.12 points. The large improvement indicates that explicitly storing reusable adaptation subspaces in the shared bases provides a substantially more structured and effective parameterization than generating the full factors directly. The complete C-MAGE configuration additionally incorporates a learned embedding associated with the corresponding department:

$$c_i = \text{sg}(z_i) + e_{d_i}^{D}.$$

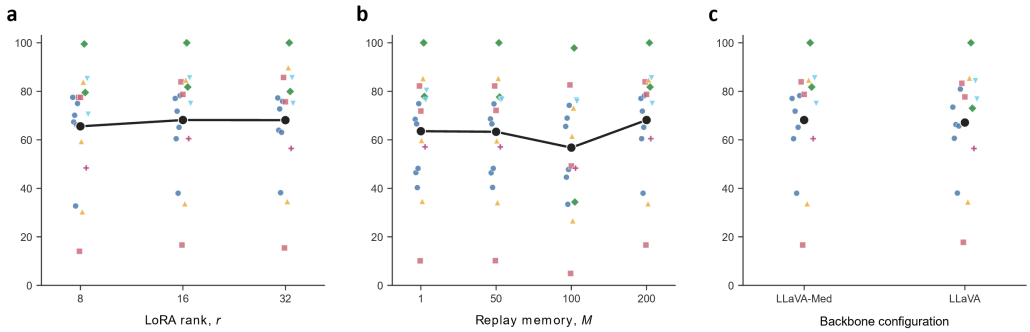


Figure 5: Sensitivity analysis of C-MAGE. (a) Sensitivity to LoRA rank. (b) Sensitivity to replay-memory size. (c) Sensitivity to the vision-language backbone.

This department-aware condition further improves the average score from 65.87 to 68.16, corresponding to a gain of 2.29 points. The department embedding therefore provides complementary clinical context that helps specialize the generated adaptations while preserving instance-level conditioning. Collectively, these results support our structured generative LoRA parameterization: shared adaptation bases enable parameter reuse and instance-specific updates within a common adaptation space, while observed department context guides clinical specialization.

D.2 Task-Order Robustness

Fig. 4(b) presents the dataset-level results under one standard and two shuffled task sequences. The sequences are defined over six department blocks: Radiology (Rad), Pathology (Path), Dermatology (Derm), Ophthalmology (Oph), Gynecology (Gyn), and Gastroenterology (Gas). The dataset order within each department is kept fixed, while the department blocks are arranged as follows:

- **Standard order:** Rad $\rightarrow$ Path $\rightarrow$ Derm $\rightarrow$ Oph $\rightarrow$ Gyn $\rightarrow$ Gas.
- **Shuffled order I:** Rad $\rightarrow$ Derm $\rightarrow$ Path $\rightarrow$ Oph $\rightarrow$ Gas $\rightarrow$ Gyn.
- **Shuffled order II:** Rad $\rightarrow$ Path $\rightarrow$ Gyn $\rightarrow$ Gas $\rightarrow$ Oph $\rightarrow$ Derm.

As shown in Fig. 4(b), C-MAGE achieves final Avg. Scores of 68.16, 66.19, and 65.52 under the standard order, Shuffled order I, and Shuffled order II, respectively. The corresponding reductions relative to the standard order are 1.97 and 2.64 points. Despite these differences, the dataset-level profiles remain broadly consistent, indicating that C-MAGE does not depend strongly on a single training curriculum. The remaining variation suggests that departments encountered earlier can influence the knowledge accumulated by the parameter generator and subsequently transferred to later tasks. All three radar charts use the same dataset-axis arrangement to facilitate direct comparison. Therefore, the clockwise arrangement of datasets in Fig. 4(b) does not represent the training order; the actual sequences are defined above.

D.3 Additional Sensitivity Experiments

Fig. 5(a)-(c) reports the sensitivity of C-MAGE to the LoRA rank, replay-memory size, and vision-language backbone. Colored markers report individual dataset scores, whereas black markers and connecting lines, where applicable, indicate the average across datasets.

LoRA rank. Fig. 5(a) compares LoRA ranks $r \in \{8, 16, 32\}$. The corresponding average final Avg. Scores are 65.56, 68.16, and 68.12. For $r = 8$ and $r = 16$, the shared-basis dimension is fixed at 128. For $r = 32$, we increase the basis dimension to 256 to provide sufficient capacity for the higher-rank adaptation. Increasing the rank from 8 to 16 improves performance by 2.60 points, whereas the higher-capacity $r = 32$ configuration provides no further improvement despite its larger rank and basis dimension. We therefore adopt $r = 16$ with a basis dimension of 128 as the default configuration, as it offers the best balance between adaptation capacity, parameter efficiency, and final-task performance.

Replay-memory size. Fig. 5(b) evaluates per-dataset replay-memory quotas $M \in \{1, 50, 100, 200\}$. C-MAGE maintains a final Avg. Score above 60 even under the extreme $M = 1$ setting, demonstrating that the method can preserve substantial previously acquired knowledge with only one replay sample per task. Performance remains similar at $M = 50$, decreases at $M = 100$, and reaches its highest value of 68.16 at $M = 200$. The non-monotonic trend indicates that replay effectiveness depends not only on memory capacity but also on the representativeness of the retained samples; a larger memory may offer limited benefit when its contents do not adequately preserve important task knowledge. Developing importance-aware replay-sample selection is therefore a promising direction for future work. We use $M = 200$ as the default setting in the main experiments to align the replay-memory budget with all comparison methods and ensure a controlled evaluation.

Vision-language backbone. Fig. 5(c) compares LLaVA-Med and LLaVA while keeping the remaining C-MAGE configuration fixed. The two backbones achieve average final Avg. Scores of 68.16 and 67.12, respectively, corresponding to a difference of only 1.04 points. The limited degradation indicates that C-MAGE can operate with different vision-language backbones, although the medically specialized LLaVA-Med backbone provides a modest advantage.

D.4 Functional Geometry and Training Paradigms

Functional-response geometry analysis. We first examine whether C-MAGE-generated LoRA adaptations reproduce the transformations of independently optimized expert adapters. For sample i, adapted layer ℓ, and projection ρ, we compute

$$\begin{aligned} \mathbf{r}^{\text{gen}}_{i,\ell,\rho} &= \frac{\alpha}{r}\mathbf{B}_{i,\ell,\rho}\mathbf{A}_{i,\ell,\rho}\bar{\mathbf{g}}_{i,\ell}, \\ \mathbf{r}^{\text{expert}}_{i,\ell,\rho} &= \frac{\alpha}{r}\mathbf{B}^{\text{expert}}_{\tau(i),\ell,\rho}\mathbf{A}^{\text{expert}}_{\tau(i),\ell,\rho}\bar{\mathbf{g}}_{i,\ell}, \end{aligned} \tag{14}$$

where $\bar{\mathbf{g}}_{i,\ell}$ is the hidden representation entering the corresponding projection, and $\tau(i)$ denotes the continual dataset index such that sample i belongs to $\mathcal{D}_{\tau(i)}$. Each residual response represents the additive change induced by a LoRA update relative to the frozen projection output, thereby characterizing the adaptation through its effect on the input representation rather than through its factor values. The first response ($\mathbf{r}^{\text{gen}}_{i,\ell,\rho}$) is induced by the sample-specific LoRA factors generated by C-MAGE, while the second ($\mathbf{r}^{\text{expert}}_{i,\ell,\rho}$) is induced by the independently optimized expert for the corresponding dataset. For each sample, we collect the generated and expert residual responses from the query and value projections at three representative adapted layers (i.e., the earliest, middle, and latest layers). We normalize each layer-projection block, concatenate the resulting blocks into a sample-level representation, and jointly embed the generated and expert representations using PCA followed by t-SNE.

Fig. 6(a) shows six classification datasets, one from each clinical department, with 200 image-instruction pairs sampled per dataset. Both C-MAGE-generated and independently optimized expert adaptations preserve clear dataset-dependent structure, while their induced responses occupy largely separated regions of the shared embedding. This shows that C-MAGE learns structured dataset-dependent specialization without simply reproducing the functional geometry of the expert bank, suggesting that effective adaptations need not follow the geometry of independently optimized experts. The expert responses form sample-wise clouds despite fixed adapter parameters because the same expert adaptation acts on different hidden representations. Thus, panel (a) reflects variation in the induced functional responses rather than variation in the expert parameters themselves. As with any t-SNE visualization, the resulting geometry is interpreted qualitatively.

Training paradigms. Following the above analysis, we next investigate whether explicitly aligning C-MAGE with independently optimized expert adapters improves parameter generation performance. In addition to the autoregressive response-generation loss $\mathcal{L}_{\text{CE}}$, we define expert alignment loss

$$\mathcal{L}_{\text{exp}} = \lambda_{\text{fac}}\mathcal{L}_{\text{fac}} + \lambda_{\text{fun}}\mathcal{L}_{\text{fun}}, \tag{15}$$

where the expert LoRA factors are first canonicalized to remove the gauge ambiguity of the factorization. Specifically, for the effective update of the expert associated with dataset t, $\Delta\mathbf{W}^{\text{exp}}_{t,\ell,\rho} = \mathbf{U}_{t,\ell,\rho}\mathbf{\Sigma}_{t,\ell,\rho}\mathbf{V}^{\top}_{t,\ell,\rho}$, we construct scale-adjusted balanced factors as $\widehat{\mathbf{B}}^{\text{exp}}_{t,\ell,\rho} = \mathbf{U}_{t,\ell,\rho}\mathbf{\Sigma}^{1/2}_{t,\ell,\rho}/\sqrt{s}$ and

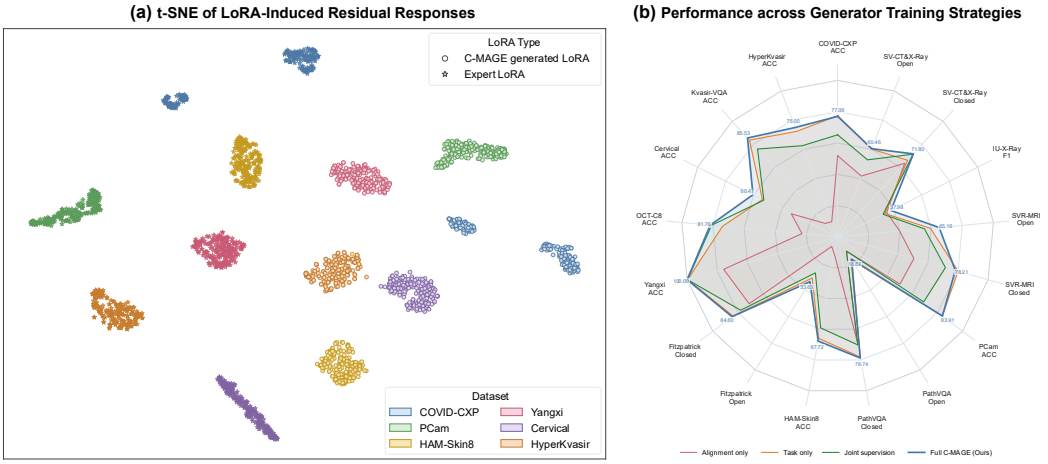


Figure 6: Training paradigms and functional geometry of generated LoRA. **(a)** Joint PCA-t-SNE visualization of the functional residual responses induced by C-MAGE-generated and independently trained expert LoRAs. Colors identify datasets, circles denote C-MAGE-generated LoRAs, and stars denote independently trained expert LoRAs. **(b)** Per-dataset comparison of training paradigms (alignment-only, task-only, joint task-and-expert alignment, and full C-MAGE training) across all evaluation metrics.

$\widehat{\mathbf{A}}^{\mathrm{exp}}_{t,\ell,\rho} = \mathbf{\Sigma}^{1/2}_{t,\ell,\rho}\mathbf{V}^{\top}_{t,\ell,\rho}/\sqrt{s}$, where $s = \alpha/r$. Consequently, $s\widehat{\mathbf{B}}^{\mathrm{exp}}_{t,\ell,\rho}\widehat{\mathbf{A}}^{\mathrm{exp}}_{t,\ell,\rho} = \Delta\mathbf{W}^{\mathrm{exp}}_{t,\ell,\rho}$. Let $\mathcal{J}$ denote the set of supervised sample-layer-projection tuples (i,ℓ,ρ), and let $\mathcal{D}(i)$ denote the dataset associated with sample i. $\mathbf{A}_{i,\ell,\rho}$ and $\mathbf{B}_{i,\ell,\rho}$ are the C-MAGE-generated LoRA factors obtained by expanding the coefficient matrices through the shared bases. The factor-alignment loss is

$$\mathcal{L}_{\mathrm{fac}} = \frac{1}{|\mathcal{J}|}\sum_{(i,\ell,\rho)\in\mathcal{J}} \frac{\mathrm{MSE}\Big(\mathbf{A}_{i,\ell,\rho}, \widehat{\mathbf{A}}^{\mathrm{exp}}_{\mathcal{D}(i),\ell,\rho}\Big) + \mathrm{MSE}\Big(\mathbf{B}_{i,\ell,\rho}, \widehat{\mathbf{B}}^{\mathrm{exp}}_{\mathcal{D}(i),\ell,\rho}\Big)}{\mathrm{MS}\Big(\widehat{\mathbf{A}}^{\mathrm{exp}}_{\mathcal{D}(i),\ell,\rho}\Big) + \mathrm{MS}\Big(\widehat{\mathbf{B}}^{\mathrm{exp}}_{\mathcal{D}(i),\ell,\rho}\Big) + \epsilon}, \tag{16}$$

where $\mathrm{MS}(\cdot)$ denotes mean-squared magnitude. To compare the induced transformations independently of the particular factorization, we further define

$$\mathcal{L}_{\mathrm{fun}} = \frac{1}{|\mathcal{J}|}\sum_{(i,\ell,\rho)\in\mathcal{J}} \frac{\mathrm{MSE}\Big(s\mathbf{B}_{i,\ell,\rho}\mathbf{A}_{i,\ell,\rho}\mathbf{Z},\ s\widehat{\mathbf{B}}^{\mathrm{exp}}_{\mathcal{D}(i),\ell,\rho}\widehat{\mathbf{A}}^{\mathrm{exp}}_{\mathcal{D}(i),\ell,\rho}\mathbf{Z}\Big)}{\mathrm{MS}\Big(s\widehat{\mathbf{B}}^{\mathrm{exp}}_{\mathcal{D}(i),\ell,\rho}\widehat{\mathbf{A}}^{\mathrm{exp}}_{\mathcal{D}(i),\ell,\rho}\mathbf{Z}\Big) + \epsilon}, \tag{17}$$

where $\mathbf{Z}$ is a fixed low-dimensional probe matrix.

We evaluate **four** training paradigms: ***alignment only***, which minimizes $\mathcal{L}_{\mathrm{exp}}$ without response supervision; ***task only***, which minimizes $\mathcal{L}_{\mathrm{CE}}$; ***joint supervision***, which combines $\mathcal{L}_{\mathrm{CE}}$ and $\mathcal{L}_{\mathrm{exp}}$; and the ***complete C-MAGE*** configuration, which uses direct task supervision together with the department prior. All variants use the same continual sequence, replay budget, backbone, and LoRA configurations. Fig. 6(b) compares the four training paradigms across all task-specific metrics. The alignment-only variant tests whether reproducing the expert transformations is sufficient without response-level supervision. Its weaker downstream performance shows that factor and functional alignment alone cannot replace the task objective. Joint task and expert-alignment supervision also generally underperforms task-only optimization, indicating that closer agreement with independently optimized adapters does not necessarily improve response generation. Since each expert is optimized independently for a single dataset, constraining the shared continual generator toward these expert solutions may unnecessarily restrict the adaptation space. In contrast, direct task-driven optimization allows C-MAGE to discover adaptations that are better aligned with the downstream response-generation objective. The complete department-conditioned C-MAGE configuration achieves the best or near-best performance on most metrics, supporting task-driven optimization guided by the department priors.

E LIMITATIONS AND SOCIETAL IMPACTS

C-MAGE remains subject to several limitations. First, inference assumes that the clinical department associated with each input is known and belongs to the set observed during training. Although the model remains competitive without department conditioning, robustness to missing, incorrect, or previously unseen department metadata requires further study. Second, C-MAGE reduces the growth of persistent adaptation parameters but still relies on replay; the overall storage requirement therefore depends on the adopted replay-memory policy. Third, the method does not completely eliminate forgetting, as reflected by its remaining negative backward transfer. Fourth, because the MLLM backbone is frozen, generated LoRA updates cannot fully compensate for visual evidence or clinical concepts that are absent from the pretrained representation. Free-form responses may additionally omit relevant findings or introduce unsupported content.

Our evaluation is based on curated public datasets and does not establish clinical validity across institutions, patient populations, acquisition protocols, or previously unseen clinical departments. Further evaluation on volumetric, longitudinal, and externally collected data, together with uncertainty estimation, calibration, and prospective clinician assessment, is necessary before clinical use. While parameter-efficient continual adaptation may reduce computational and storage requirements, erroneous or overconfident outputs could cause harm if used without appropriate human oversight.