CORE: COLLABORATIVE ORTHOGONAL RESIDUAL EXPERTS FOR MULTIMODAL CONTINUAL INSTRUCTION TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Continual Instruction Tuning (MCIT) aims to adapt multimodal large language models to sequentially arriving instruction tasks while preserving previously acquired capabilities. Recent methods address this problem by expanding task-specific LoRA experts and freezing historical experts, thereby reducing direct parameter interference. However, parameter isolation alone does not ensure effective use of newly introduced capacity: when tasks overlap, a new expert may redundantly learn representations already captured by relevant historical experts. We propose Collaborative Orthogonal Residual Experts (CORE), which reframes the newly allocated expert as a residual expert that learns with respect to reusable historical knowledge. CORE uses sample-level dual-modal reference scores to reuse relevant historical experts and select a reference expert for residual learning. Because multimodal hidden states entangle task-dependent visual evidence with reusable non-visual capabilities, CORE introduces Attention-Masked Modality Decoupling (AMMD) to construct visual-oriented representations from image-centric attention heads. Orthogonal Subspace Exploration (OSE) then discourages redundant alignment between the current and reference experts in this space, with its strength adapted by the selected reference score. Experiments on the CoIN benchmark demonstrate that CORE consistently outperforms existing methods in current-task adaptation and previous-task retention.

1 INTRODUCTION

Multimodal large language models (MLLMs) have demonstrated strong capabilities in understanding and reasoning over visual and textual information (Liu et al., 2023). Through multimodal instruction tuning, MLLMs can acquire diverse vision-language capabilities, such as visual question answering, image captioning, and visual grounding (Dai et al., 2023; Li et al., 2023; Peng et al., 2024). In real-world deployment, however, new tasks and instruction-following requirements may continue to emerge after the initial tuning stage. Retaining all historical task data and repeatedly retraining or jointly tuning on the accumulated data is computationally expensive and may be infeasible when previous data are no longer accessible. Multimodal Continual Instruction Tuning (MCIT) addresses this setting by enabling an MLLM to sequentially acquire new capabilities while preserving previously learned knowledge (Chen et al., 2024). Among existing MCIT approaches, expanding Mixture of LoRA Experts (MoLE) architectures have emerged as a promising direction. By allocating new LoRA experts to incoming tasks while freezing previously learned experts, these methods isolate task-specific updates across experts, thereby reducing direct parameter interference and mitigating catastrophic forgetting (Yu et al., 2025; Zhao et al., 2026).

However, parameter isolation alone does not ensure that newly introduced expert capacity is used effectively. Existing expert expansion methods primarily determine where task-specific knowledge is allocated, but provide limited guidance on what the newly allocated expert should learn relative to the existing expert pool (Yu et al., 2025). In MCIT, different tasks may share reusable knowledge while still requiring task-specific capabilities to solve their respective instructions. When a new task is related to previously learned tasks, historical experts may already encode representations that are useful for the current task. Without explicitly accounting for such overlap, the newly allocated expert may learn representations that substantially overlap with those already captured by relevant historical

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.

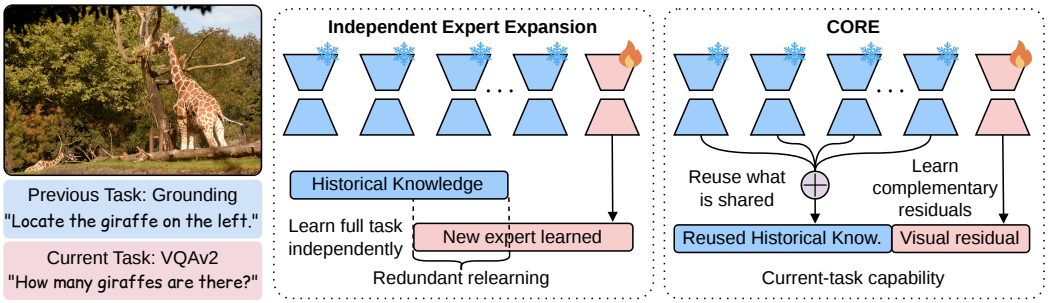


Figure 1: Visually related samples across different instruction tasks may share reusable visual knowledge while requiring different task-specific evidence. Independent expert expansion trains a newly allocated expert on the entire incoming task, which may redundantly relearn knowledge already captured by historical experts. In contrast, CORE reuses relevant historical knowledge and guides the current expert toward complementary visual residuals.

experts, resulting in redundant expert learning (Figure 1). Such redundancy limits the effective use of expanded expert capacity and weakens the specialization of newly introduced experts.

To address redundant expert learning, we reframe the newly allocated expert as a residual expert rather than an isolated task learner. The new expert should not model the entire incoming task independently; instead, it should complement the reusable representations provided by relevant historical experts. Realizing this residual-learning view in MCIT raises three challenges. First, historical knowledge reuse and reference expert selection should be sample-specific, since different samples from the same task may relate to different previous tasks. Second, the new expert requires a learning signal that discourages redundant overlap with the reference expert while retaining the ability to acquire useful current-task representations. Third, residual expert learning should account for the multimodal structure of MLLMs. Their hidden states entangle task-dependent visual evidence with reusable instruction-following, language-generation, and reasoning capabilities across tasks.

To address these challenges, we propose **Collaborative Orthogonal Residual Experts (CORE)**, a framework that guides the newly allocated LoRA expert to learn residual representations with respect to relevant historical experts. CORE assigns each historical expert a dual-modal reference anchor, consisting of image and instruction prototypes computed by averaging frozen CLIP embeddings over the task on which the expert was learned (Guo et al., 2025). For each current sample, CORE compares its image and instruction embeddings with these anchors to obtain reference scores. The normalized reference scores serve as allocation weights for reusing historical experts during current-task learning, while the highest-scoring historical expert is selected as the reference expert. Before orthogonal exploration, CORE introduces Attention-Masked Modality Decoupling (AMMD) to construct a visual-oriented representation. Motivated by the observation that attention heads differ in their roles in visual-text interaction (Kang et al., 2025), AMMD retains the outputs of image-centric heads associated with the selected reference expert and masks the remaining head outputs, focusing subsequent orthogonal exploration on visual-oriented representations. Within this space, CORE applies Orthogonal Subspace Exploration (OSE), which penalizes alignment between the current and reference experts to reduce redundant representation learning, while the instruction-tuning objective maintains current-task usefulness. Finally, the OSE objective is weighted by the selected reference score, strengthening residual exploration when the current sample is highly related to the reference expert and relaxing the constraint when their relevance is weak.

Our contributions are threefold. First, we identify redundant expert learning as an overlooked limitation of expert expansion in MCIT and reframe the newly allocated expert as a residual expert that learns with respect to reusable historical knowledge, providing a new perspective on using expanded expert capacity. Second, we propose CORE, a reference-guided residual expert learning framework that treats historical experts as both reusable knowledge sources and reference spaces. Through sample-level dual-modal reference scores and Orthogonal Subspace Exploration, CORE reuses relevant historical experts while discouraging the current expert from redundantly learning representations already captured by the selected reference expert. Third, we introduce Attention-Masked Modality Decoupling, which constructs visual-oriented representations through image-centric at-

tention heads, allowing orthogonal exploration to focus on task-dependent visual components rather than the full multimodal hidden space. Extensive experiments on the CoIN benchmark (Chen et al., 2024) demonstrate that CORE improves current-task adaptation and previous-task retention over existing methods, with comprehensive analyses supporting the effectiveness of its key designs.

2 Related Work

Multimodal Continual Instruction Tuning. Visual instruction tuning enables multimodal large language models to follow diverse image-text instructions through multimodal instruction-response training (Liu et al., 2023; Dai et al., 2023). However, most visual instruction-tuning methods assume a static setting in which all training data are jointly available, whereas deployed models may encounter new tasks sequentially without access to previous training data. Multimodal Continual Instruction Tuning (MCIT) extends instruction tuning to this setting, aiming to acquire new multimodal capabilities while retaining previously acquired knowledge (Chen et al., 2024). Existing approaches employ continual learning strategies such as rehearsal and regularization to mitigate catastrophic forgetting (Marouf et al., 2025; Qiao et al., 2025), while parameter-efficient methods isolate task-specific adaptation within lightweight modules to reduce direct parameter interference. CoIN further establishes a systematic MCIT benchmark and introduces MoELoRA, which uses modular LoRA experts for continual adaptation (Chen et al., 2024).

Mixture of LoRA Experts for MCIT. Mixture of LoRA Experts (MoLE) provides a modular approach to MCIT by combining parameter-efficient LoRA adaptation with expert routing. CoIN introduces MoELoRA, which uses a fixed pool of LoRA experts with routing for continual adaptation (Chen et al., 2024). SMoLoRA separates routing for visual understanding and instruction following (Wang et al., 2025), while D-MoLE dynamically allocates experts across transformer layers and employs an inter-modal curriculum (Ge et al., 2025). ProgLoRA progressively expands a frozen LoRA pool and reuses prior knowledge through task-aware allocation (Yu et al., 2025); HiDe-LLaVA selects task-specific LoRA experts using image-text anchors (Guo et al., 2025); and DyMoE addresses routing drift during dynamic expert expansion through token assignment and routing regularization (Zhao et al., 2026). O-LoRA reduces cross-task interference by enforcing orthogonality between task-specific LoRA update subspaces (Wang et al., 2023). Existing methods therefore primarily address expert routing, parameter isolation, dynamic allocation, expansion, historical reuse, or orthogonality-based interference reduction, while existing orthogonality-based formulations do not explicitly account for modality-entangled representations in MLLMs. In contrast, CORE treats relevant historical experts as both reusable knowledge sources and reference spaces for residual expert learning, with orthogonal exploration localized to visual-oriented representations.

3 CORE: Collaborative Orthogonal Residual Experts

This section presents **Collaborative Orthogonal Residual Experts (CORE)**, which learns each newly introduced LoRA expert as a residual learner with respect to reusable historical experts. Figure 2 provides an overview of CORE, and Appendix B gives the full algorithm. We first establish the MCIT setting and dynamically expanding LoRA expert pool, then introduce sample-level historical expert reuse and reference selection, followed by Attention-Masked Modality Decoupling, Orthogonal Subspace Exploration, and the overall training and inference procedure.

3.1 Preliminary

Given a pre-trained multimodal large language model f, Multimodal Continual Instruction Tuning (MCIT) (Chen et al., 2024) sequentially adapts f to T multimodal instruction tasks $\{\mathcal{D}_t\}_{t=1}^{T}$, where $\mathcal{D}_t = \{x_i^t\}_{i=1}^{N_t}$ and $x_i^t = (x_{\text{img},i}^t, x_{\text{ins},i}^t, x_{\text{ans},i}^t)$ consists of an image input, textual instruction, and target response. At training stage t, only the current dataset $\mathcal{D}_t$ is available, while previous task data are inaccessible. The objective is to acquire current-task knowledge while retaining knowledge learned from previous tasks. Given the image and instruction, the model is optimized to generate the target response using the standard instruction-tuning objective

$$\mathcal{L}_{\text{IT}}^t = -\frac{1}{N_t}\sum_{i=1}^{N_t} \log p_\theta \left(x_{\text{ans},i}^t \mid x_{\text{img},i}^t, x_{\text{ins},i}^t\right), \tag{1}$$

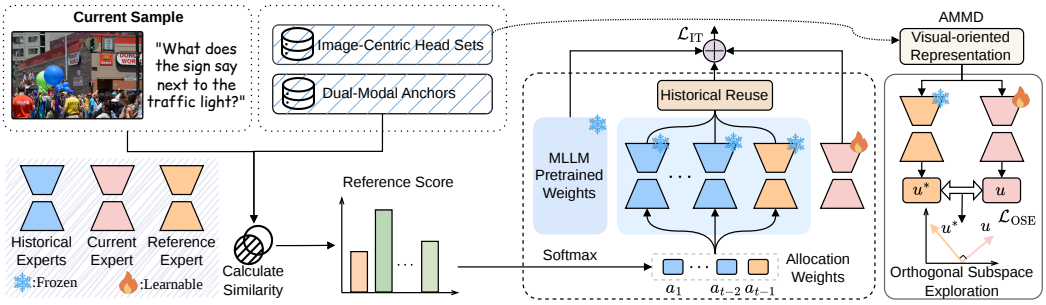


Figure 2: **Overview of the proposed CORE framework.** For each current sample, CORE computes dual-modal reference scores against the stored anchors of historical LoRA experts. The normalized scores serve as allocation weights for sample-adaptive historical reuse, while the highest-scoring expert is selected as the reference expert. AMMD uses the selected reference expert's image-centric head set to construct a visual-oriented representation. In this space, OSE penalizes alignment between the current and reference expert outputs, with its strength adapted by the selected reference score, thereby discouraging redundant representation learning. The allocation-weighted historical expert output is combined with the current expert during instruction tuning.

where θ denotes the parameters involved in adaptation, and the conditional probability follows the standard autoregressive next-token prediction objective.

3.2 Dynamic LoRA Experts

CORE parameterizes continual adaptation with dynamically expanding LoRA experts. For a frozen pre-trained linear transformation $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, the LoRA expert introduced for task t is denoted by E_t and consists of two trainable low-rank matrices $A_t \in \mathbb{R}^{r \times d_{\text{in}}}$ and $B_t \in \mathbb{R}^{d_{\text{out}} \times r}$, where $r \ll \min(d_{\text{in}}, d_{\text{out}})$. Given a hidden representation $\mathbf{h}$, the low-rank update produced by E_t is

$$E_t(\mathbf{h}) = B_t A_t \mathbf{h}. \tag{2}$$

Before learning task t, the model maintains a historical expert pool $\mathcal{E}_{t-1} = \{E_1, E_2, \ldots, E_{t-1}\}$. When task t arrives, CORE introduces the current-task expert E_t, freezes all historical experts, and optimizes only E_t. After task t is learned, E_t is frozen and added to the expert pool:

$$\mathcal{E}_t = \mathcal{E}_{t-1} \cup \{E_t\} = \{E_1, E_2, \ldots, E_t\}. \tag{3}$$

3.3 Sample-Level Historical Expert Reuse and Reference Selection

Residual expert learning first requires identifying which historical knowledge is relevant to each current sample. In MCIT, samples within the same current task may relate differently to previous tasks, making a fixed task-level reuse pattern inadequate. CORE therefore estimates the sample-level relevance between each current sample and every historical expert. Since previous task data are unavailable during subsequent training stages, each historical expert is represented by a stored dual-modal reference anchor that summarizes the image and instruction semantics of its corresponding task. Comparing a current sample with these anchors yields reference scores without accessing historical data. CORE uses these scores in two complementary ways: their normalized values serve as allocation weights for soft reuse across the historical expert pool, while the highest-scoring expert is selected as a single reference for subsequent modality decoupling and residual exploration.

Dual-Modal Reference Anchors. For each historical expert E_e, where $e < t$, CORE stores a dual-modal reference anchor after training on $\mathcal{D}_e$. CORE uses frozen CLIP image and text encoders, denoted by $f^{\text{img}}_{\text{CLIP}}$ and $f^{\text{text}}_{\text{CLIP}}$, respectively. Through contrastive image-text pretraining, CLIP learns semantically aligned visual and textual representations (Radford et al., 2021), which we use for relevance estimation without introducing additional trainable parameters. For each sample $x^e_i \in \mathcal{D}_e$, the image embedding $\mathbf{z}^e_{\text{img},i}$ and instruction embedding $\mathbf{z}^e_{\text{ins},i}$ are computed as

$$\mathbf{z}^e_{\text{img},i} = f^{\text{img}}_{\text{CLIP}}\left(x^e_{\text{img},i}\right), \qquad \mathbf{z}^e_{\text{ins},i} = f^{\text{text}}_{\text{CLIP}}\left(x^e_{\text{ins},i}\right). \tag{4}$$

CORE summarizes each modality using a mean prototype. The image anchor $\mathbf{k}^e_{\text{img}}$ and instruction anchor $\mathbf{k}^e_{\text{ins}}$ are obtained by averaging the corresponding embeddings over $\mathcal{D}_e$:

$$\mathbf{k}^e_{\text{img}} = \frac{1}{N_e}\sum_{i=1}^{N_e}\mathbf{z}^e_{\text{img},i}, \qquad \mathbf{k}^e_{\text{ins}} = \frac{1}{N_e}\sum_{i=1}^{N_e}\mathbf{z}^e_{\text{ins},i}. \tag{5}$$

Together, $\mathbf{k}^e_{\text{img}}$ and $\mathbf{k}^e_{\text{ins}}$ form the dual-modal reference anchor for E_e, requiring only two stored prototype vectors without retaining historical samples.

Sample-Level Reference Scores. At the training stage t, CORE estimates the relevance of each current sample $x^t_i \in \mathcal{D}_t$ to every historical expert E_e, where $e < t$. The current sample is embedded using the same frozen CLIP encoders, producing the image embedding $\mathbf{z}^t_{\text{img},i}$ and instruction embedding $\mathbf{z}^t_{\text{ins},i}$. Because historical relevance may depend on both visual content and instruction semantics, CORE combines the modality-specific cosine similarities as

$$s^t_{i,e} = \frac{1}{2}\left[\text{sim}\left(\mathbf{z}^t_{\text{img},i}, \mathbf{k}^e_{\text{img}}\right) + \text{sim}\left(\mathbf{z}^t_{\text{ins},i}, \mathbf{k}^e_{\text{ins}}\right)\right], \qquad e = 1, \ldots, t-1, \tag{6}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity. The resulting scalar $s^t_{i,e}$ is the reference score, with larger values indicating greater dual-modal semantic relevance to the historical task associated with E_e.

Weighted Historical Expert Reuse. CORE converts the sample-level reference scores into allocation weights to enable sample-adaptive soft reuse over the historical expert pool. Specifically, the reference scores are normalized by a softmax over historical experts:

$$a^t_{i,e} = \frac{\exp\left(s^t_{i,e}\right)}{\sum_{j=1}^{t-1}\exp\left(s^t_{i,j}\right)}, \qquad e = 1, \ldots, t-1. \tag{7}$$

Here, $a^t_{i,e}$ denotes the allocation weight of historical expert E_e for sample x^t_i, with $\sum_{e=1}^{t-1} a^t_{i,e} = 1$. Thus, historical experts contribute according to their relative relevance to each current sample.

Let $\mathbf{h}^t_i$ denote the hidden representation of x^t_i at a LoRA-adapted module, with token and layer indices omitted for simplicity. The allocation-weighted contribution from historical experts is

$$E^t_{\text{hist}}\left(\mathbf{h}^t_i\right) = \sum_{e=1}^{t-1} a^t_{i,e} E_e\left(\mathbf{h}^t_i\right). \tag{8}$$

This historical contribution is then combined with the current-task expert E_t:

$$E^t_{\text{comb}}\left(\mathbf{h}^t_i\right) = E_t\left(\mathbf{h}^t_i\right) + E^t_{\text{hist}}\left(\mathbf{h}^t_i\right). \tag{9}$$

Under this decomposition, E^t_{hist} provides reusable historical knowledge, while E_t serves as the residual component for current-task adaptation.

Reference Expert Selection. CORE uses the reference scores to select a single reference expert for each current sample. Multiple reference experts offer limited additional guidance when their representation directions are similar, whereas distinct directions may impose unnecessarily broad constraints on residual exploration. CORE therefore selects the highest-scoring historical expert:

$$e^{\star,t}_i = \arg\max_{e\in\{1,\ldots,t-1\}} s^t_{i,e}. \tag{10}$$

The corresponding reference expert and its reference score are denoted by

$$E^{\star,t}_i = E_{e^{\star,t}_i}, \qquad s^{\star,t}_i = s^t_{i,e^{\star,t}_i}. \tag{11}$$

The selected reference expert $E^{\star,t}_i$ serves as the historical reference for subsequent modality decoupling and residual exploration, while $s^{\star,t}_i$ is later used to adjust the strength of OSE.

3.4 Attention-Masked Modality Decoupling

MLLM hidden representations entangle task-dependent visual evidence with reusable non-visual information across instruction tasks. Applying OSE directly to the full hidden space may therefore

impose unnecessary orthogonal constraints on components that do not require differentiation. To localize residual exploration, CORE introduces **Attention-Masked Modality Decoupling (AMMD)**, which constructs a visual-oriented representation before OSE. AMMD identifies image-centric attention heads associated with each historical expert and uses the head set of the selected reference expert to focus subsequent residual exploration on visual-oriented components.

Image-Centric Head Bank. Since visual attention patterns vary across tasks, AMMD maintains a task-specific image-centric head set rather than a fixed set shared across tasks. Building on prior visual attention sink analysis (Kang et al., 2025), we first exclude heads whose total attention to visual tokens is below 0.2. For a sample $x_i^e \in \mathcal{D}_e$, let $\mathcal{V}_i^e$ denote the visual token positions, $\mathcal{S}_i^{e,\ell} \subseteq \mathcal{V}_i^e$ the visual sink token positions at layer ℓ, and $\mathcal{Q}_i^e$ the textual query token positions. We define the task-level visual non-sink score as

$$R_e^{\ell,h} = \frac{1}{N_e} \sum_{x_i^e \in \mathcal{D}_e} \frac{1}{|\mathcal{Q}_i^e|} \sum_{q \in \mathcal{Q}_i^e} \frac{\sum_{p \in \mathcal{V}_i^e \setminus \mathcal{S}_i^{e,\ell}} \alpha_{i,q,p}^{e,\ell,h}}{\sum_{p \in \mathcal{V}_i^e} \alpha_{i,q,p}^{e,\ell,h}}, \tag{12}$$

where $\alpha_{i,q,p}^{e,\ell,h}$ denotes the attention weight from textual query token q to visual token p. A higher $R_e^{\ell,h}$ indicates stronger task-level attention to visual non-sink tokens within task e. Among the remaining heads, the task-specific image-centric head set is then defined as

$$\mathcal{H}_e = \left\{ (\ell, h) : R_e^{\ell,h} \geq \rho \right\}, \tag{13}$$

where ρ is the head-selection threshold. Before task t, the image-centric head bank is $\{\mathcal{H}_e\}_{e=1}^{t-1}$.

Reference-Guided Head Masking. For a current sample x_i^t, AMMD retrieves the image-centric head set corresponding to the selected reference expert index $e_i^{\star,t}$:

$$\mathcal{H}_i^{\star,t} = \mathcal{H}_{e_i^{\star,t}}. \tag{14}$$

The retrieved set $\mathcal{H}_i^{\star,t}$ specifies the reference-guided image-centric heads for x_i^t. AMMD then constructs a binary mask over attention heads at each layer:

$$m_i^{\star,t,\ell,h} = \begin{cases} 1, & (\ell, h) \in \mathcal{H}_i^{\star,t}, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Masked Visual-Oriented Representation. Let M denote the number of attention heads and d_h the hidden dimension of each head. For a current sample x_i^t, the head-decomposed representation at layer ℓ before the output projection is written as

$$o_i^{t,\ell} = \left[o_i^{t,\ell,1}; o_i^{t,\ell,2}; \ldots; o_i^{t,\ell,M} \right], \tag{16}$$

where $o_i^{t,\ell,h} \in \mathbb{R}^{d_h}$ denotes the output of head h. AMMD applies the reference-guided mask to construct the visual-oriented representation as

$$v_i^{t,\ell} = \mathrm{Concat}_{h=1}^{M} \left(m_i^{\star,t,\ell,h} o_i^{t,\ell,h} \right). \tag{17}$$

By retaining the head outputs corresponding to the image-centric heads associated with the selected reference expert and suppressing the remaining heads, $v_i^{t,\ell}$ provides the visual-oriented representation used by Orthogonal Subspace Exploration for residual expert learning.

3.5 Orthogonal Subspace Exploration

Building on the visual-oriented representation $v_i^{t,\ell}$, CORE applies **Orthogonal Subspace Exploration (OSE)** between the current-task expert E_t and the selected reference expert $E_i^{\star,t}$. OSE operationalizes residual expert learning by penalizing alignment between the visual-oriented outputs of the current-task expert and the selected reference expert, discouraging the current expert from redundantly learning representations already captured by the selected reference expert.

Visual-Oriented Expert Outputs. For sample x_i^t at layer ℓ, CORE applies the current-task expert and the selected reference expert to the same visual-oriented representation:

$$u_i^{t,\ell} = E_t \left(v_i^{t,\ell} \right), \qquad u_i^{\star,t,\ell} = E_i^{\star,t} \left(v_i^{t,\ell} \right). \tag{18}$$

Here, OSE is applied specifically to the LoRA updates associated with the attention output projection of transformer layer ℓ, and $u_i^{t,\ell}$ and $u_i^{\star,t,\ell}$ denote the corresponding outputs of the current-task and selected reference experts, respectively.

Orthogonal Alignment Penalization. Given the visual-oriented expert outputs $u_i^{t,\ell}$ and $u_i^{\star,t,\ell}$, CORE defines the layer-wise orthogonal alignment penalty as

$$\ell_{\mathrm{orth}}^t(i,\ell) = \left(\frac{\left\langle u_i^{t,\ell}, u_i^{\star,t,\ell} \right\rangle}{\left(\|u_i^{t,\ell}\|_2 + \epsilon \right) \left(\|u_i^{\star,t,\ell}\|_2 + \epsilon \right)} \right)^2, \tag{19}$$

where ϵ is a small constant for numerical stability. The squared normalized inner product penalizes directional alignment regardless of sign, yielding zero for orthogonal outputs.

Reference-Score-Weighted OSE Objective. CORE adapts OSE strength using the selected reference score via the non-negative weight $\omega_i^t = \max(0, s_i^{\star,t})$. A non-positive selected reference score yields no OSE penalty. The OSE objective averages the weighted layer-wise penalties over N_t current-task samples and L transformer layers:

$$\mathcal{L}_{\mathrm{OSE}}^t = \frac{1}{N_t L} \sum_{i=1}^{N_t} \sum_{\ell=1}^{L} \omega_i^t \ell_{\mathrm{orth}}^t(i,\ell). \tag{20}$$

For $t = 1$, CORE sets $\mathcal{L}_{\mathrm{OSE}}^1 = 0$ because no historical expert is available.

3.6 Overall Training and Inference

Overall Training Objective. For task t, CORE optimizes only the current-task expert E_t using the instruction-tuning objective together with the reference-score-weighted OSE objective:

$$\mathcal{L}_{\mathrm{CORE}}^t = \mathcal{L}_{\mathrm{IT}}^t + \mathcal{L}_{\mathrm{OSE}}^t. \tag{21}$$

Here, $\mathcal{L}_{\mathrm{IT}}^t$ drives current-task adaptation, while $\mathcal{L}_{\mathrm{OSE}}^t$ discourages redundant alignment with the selected reference expert in the visual-oriented space, jointly promoting task-useful residual learning.

Inference. Unlike training, where E_t serves as the current residual expert, no expert is designated as current for a test sample at inference. CORE therefore treats all learned experts symmetrically and performs sample-adaptive allocation over the complete frozen expert pool $\mathcal{E}_t = \{E_e\}_{e=1}^t$. Given a sample $x = (x_{\mathrm{img}}, x_{\mathrm{ins}})$, CORE computes reference scores $s_e(x)$ against all stored dual-modal anchors and normalizes them into allocation weights $a_e(x) = \exp(s_e(x)) / \sum_{j=1}^t \exp(s_j(x))$, $e = 1, \ldots, t$. For a hidden representation $\mathbf{h}$, CORE forms the allocation-weighted expert output used with the frozen backbone to generate the response:

$$E_{\mathrm{inf}}^t(\mathbf{h}; x) = \sum_{e=1}^{t} a_e(x) E_e(\mathbf{h}). \tag{22}$$

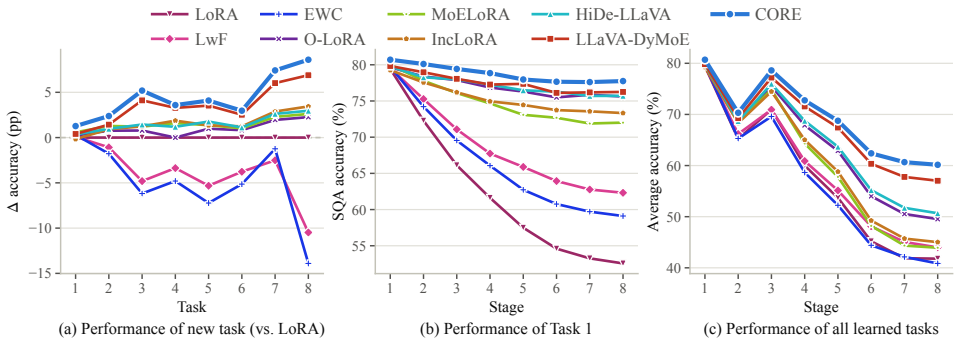
4 Experiments

4.1 Experimental Setup

We evaluate CORE on the CoIN benchmark (Chen et al., 2024) for MCIT, comprising eight sequential multimodal instruction tasks, and compare it with a diverse set of baselines, including recent expert-based MCIT methods (Zhao et al., 2026). All methods follow the same evaluation protocol with LLaVA-v1.5-7B (Liu et al., 2023) as the backbone and are evaluated using Mean Final Accuracy (MFN), Mean Average Accuracy (MAA), and Backward Transfer (BWT). Dataset statistics, baseline details, metric definitions, and implementation settings are provided in Appendix A.

4.2 Main Results

Overall Performance. Table 1 summarizes the final task-wise and aggregate performance on CoIN. CORE achieves the best MFN, MAA, and BWT, reaching 60.15%, 69.29%, and -3.83%, respectively. Compared with the strongest baseline, LLaVA-DyMoE, CORE yields absolute gains



(a) Performance of new task (vs. LoRA) (b) Performance of Task 1 (c) Performance of all learned tasks

Figure 3: Stage-wise continual learning dynamics on CoIN.

of 3.12, 1.73, and 2.21 in MFN, MAA, and BWT, respectively, while outperforming it across all eight tasks, indicating stronger current-task adaptation and previous-task retention. The largest task-wise gain occurs on REF (+5.96), followed by GQA (+4.01), VQAv2 (+3.85), VizWiz (+3.63), and TextVQA (+3.38). The substantial margin over IncLoRA supports CORE's residual-learning formulation beyond independent expert expansion.

Table 1: Overall performance on CoIN: task-wise final accuracy and aggregate metrics (%).

Method	Final Accuracy on Each Task (%)								Aggregate Metrics (%)		
	SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA^{v2}	VQA^{OCR}	MFN↑	MAA↑	BWT↑
LwF	62.32	48.66	51.45	45.84	43.76	0.24	54.96	44.63	43.98	58.72	−11.65
EWC	59.11	47.21	39.88	45.12	35.33	2.72	56.29	41.21	40.86	56.61	−13.69
LoRA	52.56	48.12	39.27	44.47	37.46	1.22	56.10	55.11	41.79	57.27	−17.75
O-LoRA	75.61	49.98	78.24	44.18	30.70	4.66	55.51	57.37	49.53	63.58	−10.98
MoELoRA	72.01	46.89	44.75	42.79	28.22	3.31	55.74	57.72	43.93	60.26	−17.11
IncLoRA	73.33	44.32	54.59	44.07	25.93	4.49	54.91	58.55	45.02	60.73	−16.12
HiDe-LLaVA	75.71	50.57	80.91	44.82	33.99	5.36	55.93	58.07	50.67	64.31	−10.42
LLaVA-DyMoE	76.25	53.86	95.80	48.40	52.35	9.25	58.30	62.00	57.03	67.56	−6.04
CORE	**77.75**	**57.24**	**96.73**	**52.41**	**55.98**	**15.21**	**62.15**	**63.71**	**60.15**	**69.29**	**−3.83**

Continual Learning Dynamics. Figure 3(a) reports improvement over LoRA on each newly learned task immediately after it is learned, thereby reducing the influence of task-difficulty differences; CORE achieves positive gains from Task 2 onward and generally the largest gains across methods, with gains persisting in later stages when more historical experts are available for reference-guided reuse. Figure 3(b) shows that CORE's Task-1 accuracy declines more slowly and remains the highest in later stages, indicating stronger long-term retention. By encouraging each new expert to complement rather than relearn reusable historical representations, CORE may reduce interference when later experts are combined with historical experts at inference. Figure 3(c) further shows that CORE maintains the highest average accuracy over all tasks learned up to each stage, with its margin over competing methods becoming more pronounced in later stages.

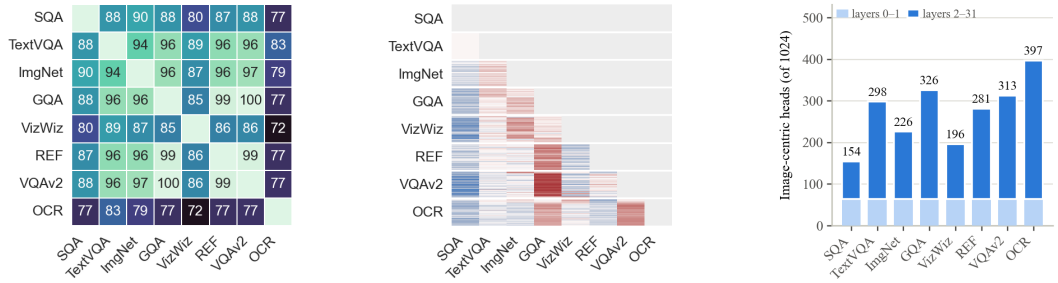
4.3 Ablation and Further Analysis

Table 2 examines CORE's key components using IncLoRA as the independent expert expansion baseline. Historical expert reuse improves MFN, MAA, and BWT from 45.02%, 60.73%, and −16.12% to 55.11%, 66.85%, and −7.38%, respectively, demonstrating the benefit of historical reuse. Applying OSE directly in the full hidden space does not further improve performance, consistent with our motivation that multimodal hidden representations entangle task-dependent visual evidence with reusable non-visual capabilities. Restricting OSE to the AMMD visual-oriented space substantially improves all three metrics, indicating that OSE effectiveness depends on the representation space. Finally, reference-score-weighted OSE (CORE) further improves performance, supporting reference-score-based adaptation of OSE strength.

Table 2: Component ablation of CORE on CoIN.

Variant	Final Accuracy on Each Task (%)								Aggregate Metrics (%)		
	SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA^{v2}	VQA^{OCR}	MFN↑	MAA↑	BWT↑
IncLoRA	73.33	44.32	54.59	44.07	25.93	4.49	54.91	58.55	45.02	60.73	−16.12
+ Reuse	76.59	51.06	96.33	47.54	44.60	6.74	57.44	60.57	55.11	66.85	−7.38
Reuse + OSE_{full}	75.90	50.17	96.29	46.79	43.59	6.66	56.63	60.01	54.51	66.41	−7.55
Reuse + OSE_{AMMD}	77.10	56.16	96.52	51.49	54.60	14.16	61.14	62.86	59.25	68.88	−4.46
CORE	**77.75**	**57.24**	**96.73**	**52.41**	**55.98**	**15.21**	**62.15**	**63.71**	**60.15**	**69.29**	**−3.83**

Inter-Task Visual Overlap. Figure 4(a) shows pairwise cosine similarity between task-level image anchors, with several task pairs exhibiting substantial visual similarity. Such similarity suggests opportunities for historical reuse but also a risk of redundant relearning, motivating relevance-weighted historical reuse and reference expert selection rather than uniform treatment of historical experts. Appendix D shows that stronger historical relevance is associated with greater current/reference expert overlap under historical reuse, while CORE substantially reduces this overlap.



(a) Inter-task visual overlap. (b) Sample-level references. (c) Image-centric head counts.

Figure 4: (a) Pairwise cosine similarity between task-level image anchors. (b) Dual-modal reference scores between current-task samples and historical tasks; red cells mark the highest-scoring task whose expert is selected as the reference expert, and blue the others. (c) Image-centric head counts per CoIN task; light shading marks the first two layers, where visual sink tokens have not emerged.

Sample-Level Reference Selection. Figure 4(b) shows dual-modal reference scores between sampled current-task samples and historical tasks, with the highest-scoring task highlighted in red and its expert selected as the reference expert. Samples within the same current task can select different historical experts, indicating sample-specific historical relevance and supporting sample-level selection based on visual content and instruction semantics rather than a fixed task-level reference.

Task-Dependent Image-Centric Head Sets. Figure 4(c) reports image-centric head counts for each CoIN task. Light shading marks the first two layers, where visual sink tokens have not yet emerged and visual non-sink scores remain close to one. Across remaining layers, head counts vary by task, indicating task-dependent visual attention patterns and supporting AMMD's task-specific image-centric head bank. Complete layer-by-head visual non-sink score maps are provided in Figure 6.

5 Conclusion

This paper addresses redundant expert learning in MCIT by reframing newly introduced experts as residual experts with respect to reusable historical knowledge. CORE uses sample-level dual-modal reference scores for adaptive historical reuse and reference selection, while AMMD and OSE guide the current expert toward complementary representations in the visual-oriented space. Experiments on CoIN demonstrate that CORE improves current-task adaptation and previous-task retention over existing methods, with ablation studies and further analyses supporting the effectiveness of its key designs. Overall, these results suggest that continual expert expansion benefits from coordinating historical knowledge reuse with the complementary use of newly introduced capacity.

AI USE STATEMENT

In this work, we used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, and assist in the writing of proofs. We did not use generative AI tools to generate synthetic datasets, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean or reformat datasets, support qualitative or thematic data analysis, or interpret results. Additionally, we used generative AI tools to edit the manuscript for readability, format references, and suggest the structure of the paper. We reviewed all AI-assisted work and verified the conceptual and mathematical content. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates continual adaptation of multimodal large language models using existing benchmark datasets and pretrained models and does not involve new data collection or human-subject research. CORE focuses on continual learning performance and does not explicitly address safety, fairness, privacy, or bias. Models adapted with CORE may therefore inherit limitations and risks from the underlying pretrained models and the data used for continual adaptation. Deployment in sensitive or high-stakes settings should be accompanied by application-specific evaluation and appropriate safeguards.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 provides the formulation of CORE, while Appendix A documents the datasets, baselines, evaluation metrics, and implementation settings used in our experiments. Appendix B provides the full training and inference algorithm, and Appendix C presents the assumptions and proofs for the analytical results. Additional analyses supporting the reported results are provided in Appendix D. An anonymized code repository is available at `https://anonymous.4open.science/r/CORE-2F1D/`.

REFERENCES

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. *NeurIPS*, 37:57817–57840, 2024.

Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ip, and Sam Kwong. SEFE: Superficial and essential forgetting eliminator for multimodal continual instruction tuning. In *ICML*, 2025. URL `https://openreview.net/forum?id=teJdFzLnKh`.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 36:49250–49267, 2023.

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 *ICML*, ICML'25. JMLR.org, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp. 6904–6913, 2017.

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 *ACL*, pp. 13572–13586, 2025.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pp. 3608–3617, 2018.

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

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pp. 6700–6709, 2019.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *ICLR*, 2025. URL https://openreview.net/forum?id=7uDI7w5RQA.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pp. 787–798, 2014.

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. *PNAS*, 114(13):3521–3526, 2017.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pp. 19730–19742. PMLR, 2023.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *TPAMI*, 40(12):2935–2947, 2017.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 35:2507–2521, 2022.

Imad Eddine Marouf, Enzo Tartaglione, Stéphane Lathuilière, and Joost Van De Weijer. Ask and remember: A questions-only replay strategy for continual visual question answering. In *ICCV*, pp. 18078–18089, 2025.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pp. 947–952. IEEE, 2019.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *ICLR*, volume 2024, pp. 51575–51598, 2024.

Jingyang Qiao, Zhizhong Zhang, Xin Tan, Yanyun Qu, Shouhong Ding, and Yuan Xie. Large continual instruction assistant. In *ICML*, ICML'25. JMLR.org, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PmLR, 2021.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pp. 8317–8326, 2019.

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 *EMNLP*, pp. 10658–10671, 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 *ICCV*, pp. 177–186, 2025.

Yahan Yu, Duzhen Zhang, Yong Ren, Xuanle Zhao, Xiuyi Chen, and Chenhui Chu. Progressive lora for multimodal continual instruction tuning. In *ACL*, pp. 2779–2796, 2025.

Chongyang Zhao, Mingsong Li, Haodong Lu, and Dong Gong. On token's dilemma: Dynamic moe with drift-aware token assignment for continual learning of large vision language models. In *CVPR*, pp. 3941–3952, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *NeurIPS*, 36:46595–46623, 2023.

A EXPERIMENTAL SETUP

Datasets. We evaluate CORE on the CoIN benchmark (Chen et al., 2024), which comprises eight multimodal instruction tasks: ScienceQA (Lu et al., 2022), TextVQA (Singh et al., 2019), ImageNet (Russakovsky et al., 2015), GQA (Hudson & Manning, 2019), VizWiz (Gurari et al., 2018), Grounding (Kazemzadeh et al., 2014), VQAv2 (Goyal et al., 2017), and OCR-VQA (Mishra et al., 2019). Together, these tasks span diverse data scales, visual styles, instruction formats, and domain characteristics, providing a comprehensive testbed for Multimodal Continual Instruction Tuning. CoIN contains approximately 569k training samples and 261k test samples in total. Table 3 summarizes the task type and dataset size for each task.

Table 3: Statistics of the eight multimodal instruction tasks in the CoIN benchmark.

Task Type	**Dataset**	**Train**	**Test**
Knowledge Grounded IQA	ScienceQA	12k	4k
Reading Comprehension IQA	TextVQA	34k	5k
Classification	ImageNet	129k	5k
Visual Reasoning IQA	GQA	72k	1k
Blind People IQA	VizWiz	20k	8k
Grounding	RefCOCO RefCOCO+ RefCOCOg	55k	31k
Image Question Answering (IQA)	VQAv2	82k	107k
OCR IQA	OCR-VQA	165k	100k
Total	–	**569k**	**261k**

Baselines. We compare CORE with a diverse set of representative continual learning and parameter-efficient tuning baselines. As a basic expansion baseline, we include IncLoRA, which introduces a new LoRA module for each incoming task without historical expert reuse or residual-learning guidance. Following the IncLoRA setting in LLaVA-DyMoE (Zhao et al., 2026), previously learned LoRA modules remain frozen after their corresponding tasks are completed. At inference after task t, all learned LoRA modules up to the current stage are jointly activated, and their outputs are directly summed without task-specific routing or sample-adaptive weighting. We further compare with LwF (Li & Hoiem, 2017), EWC (Kirkpatrick et al., 2017), LoRA (Hu et al., 2022), O-LoRA (Wang et al., 2023), MoELoRA (Chen et al., 2024), ProgLoRA (Yu et al., 2025), HiDe-LLaVA (Guo et al., 2025), and LLaVA-DyMoE (Zhao et al., 2026).

Evaluation Metrics. Following CoIN (Chen et al., 2024) and subsequent MCIT studies (Yu et al., 2025; Zhao et al., 2026), we report three standard continual learning metrics: Mean Final Accuracy (MFN), Mean Average Accuracy (MAA), and Backward Transfer (BWT). Let $A_{j,i}$ denote the accuracy on task i after completing training on task j, and let T denote the total number of tasks. Specifically, MFN assesses the average accuracy across all tasks after the complete continual training sequence, computed as $\mathrm{MFN} = \frac{1}{T}\sum_{i=1}^{T} A_{T,i}$. MAA represents the mean of the average accuracies over all learned tasks after each continual training stage, computed as $\mathrm{MAA} = \frac{1}{T}\sum_{j=1}^{T}\frac{1}{j}\sum_{i=1}^{j} A_{j,i}$. BWT assesses the degree of forgetting by comparing the final performance on each task with its performance immediately after being learned, computed as $\mathrm{BWT} = \frac{1}{T}\sum_{i=1}^{T}(A_{T,i} - A_{i,i})$. Higher MFN and MAA indicate better overall performance, while higher BWT indicates less forgetting.

Implementation Details. We use LLaVA-v1.5-7B (Liu et al., 2023) as the backbone model for all experiments, with Vicuna (Zheng et al., 2023) as the language backbone and CLIP ViT-L/14 (Radford et al., 2021) as the visual encoder. During continual training, the pretrained backbone and all historical experts remain frozen, and only the newly introduced LoRA expert for the current task is optimized. For each task, we introduce a rank-4 LoRA expert into all linear layers of the language model. For reference-anchor construction and reference scoring, we use frozen CLIP image and text encoders to extract image and instruction representations. For AMMD, we set the image-centric head-selection threshold to $\rho = 0.8$ (Kang et al., 2025). We optimize the model using AdamW with a global batch size of 128, a learning rate of 2×10^{-4}, and a warmup ratio of 0.03. All experiments are conducted on two NVIDIA H200 GPUs using BF16 precision and DeepSpeed ZeRO-2.

Algorithm 1 Full Training and Inference Pipeline of CORE

Require: Sequential datasets $\{\mathcal{D}_t\}_{t=1}^{T}$; frozen pretrained MLLM f; frozen CLIP image/text encoders; head-selection threshold ρ

Ensure: Frozen expert pool $\mathcal{E}_T$, dual-modal reference anchors, and image-centric head sets

Initialize $\mathcal{E}_0 \leftarrow \varnothing$; initialize the anchor bank and image-centric head bank as empty
for $t = 1, \ldots, T$ **do**
  Initialize current-task LoRA expert E_t
  Freeze the pretrained backbone and all historical experts in $\mathcal{E}_{t-1}$
  for each minibatch $\mathcal{B} \subset \mathcal{D}_t$ **do**
    if $t > 1$ **then**
      for each sample $x_i^t \in \mathcal{B}$ **do**
        Compute current image/instruction CLIP embeddings
        Compute reference scores $\{s_{i,e}^t\}_{e=1}^{t-1}$ against all stored dual-modal anchors
        Compute allocation weights
        $a_{i,e}^t = \exp(s_{i,e}^t) / \sum_{j=1}^{t-1} \exp(s_{i,j}^t)$
        Main instruction-tuning branch:
        $E_{\text{hist}}^t(\mathbf{h}_i^t) = \sum_{e=1}^{t-1} a_{i,e}^t E_e(\mathbf{h}_i^t)$
        $E_{\text{comb}}^t(\mathbf{h}_i^t) = E_t(\mathbf{h}_i^t) + E_{\text{hist}}^t(\mathbf{h}_i^t)$
        AMMD–OSE residual-learning branch:
        Select $e_i^{\star,t} = \arg\max_{e \in \{1,\ldots,t-1\}} s_{i,e}^t$
        Set $E_i^{\star,t} = E_{e_i^{\star,t}}$ and $s_i^{\star,t} = s_{i,e_i^{\star,t}}^t$
        Retrieve $\mathcal{H}_i^{\star,t} = \mathcal{H}_{e_i^{\star,t}}$
        Apply AMMD to construct $v_i^{t,\ell}$ for each transformer layer ℓ
        Compute $u_i^{t,\ell} = E_t(v_i^{t,\ell})$ and $u_i^{\star,t,\ell} = E_i^{\star,t}(v_i^{t,\ell})$
        Compute $\ell_{\text{orth}}^t(i,\ell)$ and $\omega_i^t = \max(0, s_i^{\star,t})$
      end for
      Compute $\mathcal{L}_{\text{IT}}^t$ and aggregate $\mathcal{L}_{\text{OSE}}^t$
    else
      Compute $\mathcal{L}_{\text{IT}}^1$ using E_1 and set $\mathcal{L}_{\text{OSE}}^1 = 0$
    end if
    Compute $\mathcal{L}_{\text{CORE}}^t = \mathcal{L}_{\text{IT}}^t + \mathcal{L}_{\text{OSE}}^t$
    Update only E_t
  end for
  Freeze E_t
  Compute and store $\mathbf{k}_{\text{img}}^t$ and $\mathbf{k}_{\text{ins}}^t$ from $\mathcal{D}_t$
  Exclude heads with total visual attention below 0.2
  Compute $R_t^{\ell,h}$ and store $\mathcal{H}_t = \{(\ell,h) : R_t^{\ell,h} \geq \rho\}$
  $\mathcal{E}_t \leftarrow \mathcal{E}_{t-1} \cup \{E_t\}$
  Discard $\mathcal{D}_t$
end for

Inference after task T:

for each test sample $x = (x_{\text{img}}, x_{\text{ins}})$ **do**
  Compute reference scores $\{s_e(x)\}_{e=1}^{T}$ against all stored dual-modal anchors
  Compute $a_e(x) = \exp(s_e(x)) / \sum_{j=1}^{T} \exp(s_j(x))$
  Compute $E_{\text{inf}}^T(\mathbf{h}; x) = \sum_{e=1}^{T} a_e(x) E_e(\mathbf{h})$
  Generate the response with the frozen backbone and E_{inf}^T
end for

B FULL ALGORITHM OF CORE

Algorithm 1 summarizes the complete training and inference pipeline of CORE.

C Analytical Properties of Residual Expert Learning

This section characterizes two mathematical properties directly induced by the residual-learning design of CORE. We first analyze Orthogonal Subspace Exploration (OSE) and show that its direct output-space optimization monotonically reduces the squared normalized alignment between the current and selected reference expert outputs. We then analyze Attention-Masked Modality Decoupling (AMMD) and show that its reference-guided head masking localizes the direct OSE gradient to the selected image-centric head components. Together, these results characterize the directional and localized mechanisms underlying residual expert learning in CORE.

C.1 Orthogonal Subspace Exploration for Residual Learning

For a fixed current-task sample x_i^t and layer ℓ, let

$$u = u_i^{t,\ell}, \qquad r = u_i^{\star,t,\ell} \tag{23}$$

denote the outputs of the current expert and the selected reference expert, respectively. We analyze the direct OSE update with respect to the current expert output u, treating the nonzero selected reference expert output r as fixed.

Definition 1 (Reference-Aligned Decomposition). For nonzero r, the current expert output can be decomposed as

$$u = u_{\parallel} + u_{\perp}, \tag{24}$$

where

$$u_{\parallel} = \frac{u^\top r}{\|r\|_2^2} r, \qquad u_{\perp} = u - u_{\parallel}, \tag{25}$$

and

$$u_{\perp}^\top r = 0. \tag{26}$$

Here, $u_{\parallel}$ is the component of the current expert output along the selected reference expert direction, while $u_{\perp}$ lies in its orthogonal complement.

Lemma 1 (Relative Reference Alignment). For nonzero u and r, define the squared normalized reference alignment as

$$g(u, r) = \frac{(u^\top r)^2}{\|u\|_2^2 \|r\|_2^2}. \tag{27}$$

Then

$$g(u, r) = \frac{\|u_{\parallel}\|_2^2}{\|u\|_2^2}, \tag{28}$$

and therefore

$$0 \le g(u, r) \le 1. \tag{29}$$

Moreover, $g(u, r) = 0$ if and only if $u \perp r$, while $g(u, r) = 1$ if and only if u and r are parallel or anti-parallel.

Proof. From Definition 1,

$$u_{\parallel} = \frac{u^\top r}{\|r\|_2^2} r, \tag{30}$$

and hence

$$\|u_{\parallel}\|_2^2 = \frac{(u^\top r)^2}{\|r\|_2^2}. \tag{31}$$

Dividing by $\|u\|_2^2$ gives

$$\frac{\|u_{\parallel}\|_2^2}{\|u\|_2^2} = \frac{(u^\top r)^2}{\|u\|_2^2 \|r\|_2^2} = g(u, r). \tag{32}$$

Since $u_{\parallel}$ is the orthogonal projection of u onto the direction spanned by r,

$$0 \le \|u_{\parallel}\|_2^2 \le \|u\|_2^2, \tag{33}$$

which establishes $0 \le g(u, r) \le 1$. The equality conditions follow directly from the projection geometry.

Lemma 1 shows that squared normalized reference alignment exactly measures the relative squared magnitude of the current expert output expressed along the selected reference expert direction. Reducing this quantity therefore suppresses the reference-aligned component and promotes complementary residual directions with respect to the selected reference expert.

Proposition 1 (Monotonic Reduction of Reference Alignment under OSE). For nonzero u, a fixed nonzero reference expert output r, and $\epsilon > 0$, the ϵ-stabilized OSE penalty used by CORE is

$$\ell^{\epsilon}_{\mathrm{orth}}(u, r) = \frac{(u^\top r)^2}{(\|u\|_2 + \epsilon)^2(\|r\|_2 + \epsilon)^2}. \tag{34}$$

For a fixed non-negative OSE weight ω, consider the weighted local objective

$$\phi_\epsilon(u, r) = \omega \ell^{\epsilon}_{\mathrm{orth}}(u, r). \tag{35}$$

Under the direct output-space gradient flow

$$\frac{du}{d\tau} = -\nabla_u \phi_\epsilon(u, r), \tag{36}$$

the squared normalized reference alignment $g(u, r)$ evolves according to

$$\frac{dg}{d\tau} = -\frac{4\omega\|r\|_2^2}{(\|u\|_2 + \epsilon)^2(\|r\|_2 + \epsilon)^2} g(1 - g) \le 0. \tag{37}$$

Consequently, whenever

$$\omega > 0 \qquad \text{and} \qquad 0 < g(u, r) < 1, \tag{38}$$

the squared normalized alignment between the current and selected reference expert outputs strictly decreases along the direct OSE gradient flow.

Proof. The implemented OSE penalty can be written as

$$\ell^{\epsilon}_{\mathrm{orth}}(u, r) = q(u, r)g(u, r), \tag{39}$$

where

$$q(u, r) = \frac{\|u\|_2^2\|r\|_2^2}{(\|u\|_2 + \epsilon)^2(\|r\|_2 + \epsilon)^2}. \tag{40}$$

For fixed r, $q(u, r)$ depends on u only through its norm. Therefore, its gradient with respect to u is parallel to u.

The gradient of the squared normalized reference alignment is

$$\nabla_u g = \frac{2u^\top r}{\|u\|_2^2\|r\|_2^2}\left(r - \frac{u^\top r}{\|u\|_2^2}u\right). \tag{41}$$

Taking its inner product with u gives

$$u^\top \nabla_u g = 0. \tag{42}$$

Since $\nabla_u q$ is parallel to u,

$$\nabla_u g^\top \nabla_u q = 0. \tag{43}$$

Moreover,

$$\left\| r - \frac{u^\top r}{\|u\|_2^2}u \right\|_2^2 = \|r\|_2^2(1 - g), \tag{44}$$

which yields

$$\|\nabla_u g\|_2^2 = \frac{4}{\|u\|_2^2} g(1 - g). \tag{45}$$

Using

$$\phi_\epsilon = \omega q g, \tag{46}$$

its gradient is

$$\nabla_u \phi_\epsilon = \omega\left(q\nabla_u g + g\nabla_u q\right). \tag{47}$$

Therefore, along the direct output-space gradient flow,

$$\begin{aligned}\frac{dg}{d\tau} &= \nabla_u g^\top \frac{du}{d\tau} \\ &= -\nabla_u g^\top \nabla_u \phi_\epsilon \\ &= -\omega q \|\nabla_u g\|_2^2,\end{aligned} \tag{48}$$

where the cross term vanishes because

$$\nabla_u g^\top \nabla_u q = 0. \tag{49}$$

Substituting q and $\|\nabla_u g\|_2^2$ gives

$$\frac{dg}{d\tau} = -\frac{4\omega\|r\|_2^2}{(\|u\|_2+\epsilon)^2(\|r\|_2+\epsilon)^2} g(1-g), \tag{50}$$

which proves the proposition.

Proposition 1 establishes the directional effect induced by the actual OSE objective used in CORE at the expert-output level. For positively weighted samples, OSE monotonically decreases the relative component of the current expert output aligned with the selected reference expert whenever the two outputs are neither orthogonal nor collinear. OSE therefore provides a direct optimization mechanism for encouraging complementary residual directions relative to the selected reference expert.

Joint Objective Interpretation. At task t, CORE optimizes

$$\mathcal{L}_{\mathrm{CORE}}^t = \mathcal{L}_{\mathrm{IT}}^t + \mathcal{L}_{\mathrm{OSE}}^t. \tag{51}$$

The instruction-tuning objective drives the current expert to capture current-task requirements, while Proposition 1 shows that OSE provides a complementary directional signal that suppresses reference-aligned components. Their joint optimization therefore encourages the current expert to learn task-useful residual representations that complement reused historical expertise.

The OSE weight

$$\omega_i^t = \max(0, s_i^{\star,t}) \tag{52}$$

further makes this residual-learning signal sample-adaptive. Larger positive selected reference scores increase the contribution of the corresponding OSE term, thereby strengthening residual exploration for samples with stronger selected-reference evidence.

C.2 Attention-Masked Modality Decoupling for Localized Residual Exploration

While OSE provides the directional signal for residual exploration between expert outputs, AMMD determines which visual-oriented representation components contribute to this signal. For a fixed current-task sample x_i^t and layer ℓ, recall the head-decomposed representation

$$o_i^{t,\ell} = \left[o_i^{t,\ell,1}; \ldots; o_i^{t,\ell,M}\right], \tag{53}$$

where

$$o_i^{t,\ell,h} \in \mathbb{R}^{d_h}. \tag{54}$$

AMMD retrieves the image-centric head set $\mathcal{H}_i^{\star,t}$ associated with the selected reference expert and constructs the binary mask $m_i^{\star,t,\ell,h}$.

Reference-Guided Image-Centric Projection. Define the block-diagonal masking operator

$$P_i^{\star,t,\ell} = \mathrm{Diag}\left(m_i^{\star,t,\ell,1} I_{d_h}, \ldots, m_i^{\star,t,\ell,M} I_{d_h}\right). \tag{55}$$

Since

$$m_i^{\star,t,\ell,h} \in \{0,1\}, \tag{56}$$

the operator satisfies

$$\left(P_i^{\star,t,\ell}\right)^2 = P_i^{\star,t,\ell}, \qquad \left(P_i^{\star,t,\ell}\right)^\top = P_i^{\star,t,\ell}. \tag{57}$$

It therefore defines an orthogonal projection onto the reference-guided image-centric subspace

$$\mathcal{U}_i^{\star,t,\ell} = \mathrm{Im}\left(P_i^{\star,t,\ell}\right). \tag{58}$$

The AMMD representation in the main method can equivalently be written as

$$v_i^{t,\ell} = P_i^{\star,t,\ell} o_i^{t,\ell}, \tag{59}$$

so that

$$v_i^{t,\ell} \in \mathcal{U}_i^{\star,t,\ell}. \tag{60}$$

Proposition 2 (Gradient Localization under AMMD). For a fixed current-task sample x_i^t and layer ℓ, treat the retrieved binary head mask $P_i^{\star,t,\ell}$ as fixed and let

$$P = P_i^{\star,t,\ell}, \qquad \mathcal{U} = \mathrm{Im}(P), \qquad v = Po. \tag{61}$$

Let $\phi(v)$ denote the corresponding layer-wise reference-score-weighted OSE term computed from the visual-oriented representation v, and define

$$\widetilde{\phi}(o) = \phi(Po). \tag{62}$$

Then the direct OSE gradient with respect to the head-decomposed representation satisfies

$$\nabla_o \widetilde{\phi}(o) = P \nabla_v \phi(v) \in \mathcal{U}. \tag{63}$$

Consequently, for every masked-out attention head

$$(\ell, h) \notin \mathcal{H}_i^{\star,t}, \tag{64}$$

the corresponding direct OSE gradient is

$$\nabla_{o_i^{t,\ell,h}} \widetilde{\phi} = 0. \tag{65}$$

Thus, AMMD localizes the direct OSE signal to the reference-guided image-centric head components.

Proof. Since

$$v = Po, \tag{66}$$

the chain rule gives

$$\nabla_o \widetilde{\phi}(o) = P^\top \nabla_v \phi(v). \tag{67}$$

Because $P^\top = P$,

$$\nabla_o \widetilde{\phi}(o) = P \nabla_v \phi(v). \tag{68}$$

Therefore,

$$\nabla_o \widetilde{\phi}(o) \in \mathrm{Im}(P) = \mathcal{U}. \tag{69}$$

For head h, the corresponding block of P is

$$m_i^{\star,t,\ell,h} I_{d_h}, \tag{70}$$

so

$$\nabla_{o_i^{t,\ell,h}} \widetilde{\phi} = m_i^{\star,t,\ell,h} \nabla_{v_i^{t,\ell,h}} \phi. \tag{71}$$

For every masked-out head,

$$m_i^{\star,t,\ell,h} = 0, \tag{72}$$

which gives

$$\nabla_{o_i^{t,\ell,h}} \widetilde{\phi} = 0. \tag{73}$$

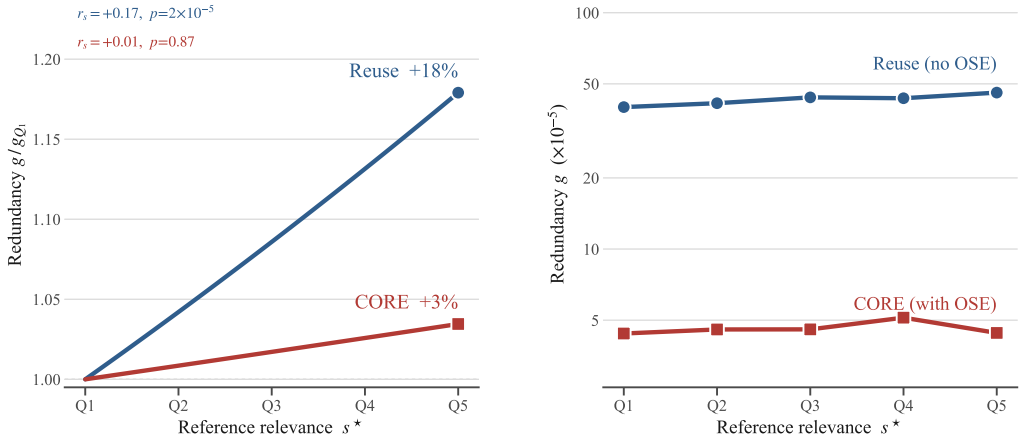
This proves the proposition.

Proposition 2 shows that AMMD provides a precise localization mechanism for OSE. The reference-guided image-centric head mask determines which visual-oriented representation components contribute directly to the OSE signal, while Proposition 1 characterizes how this signal reduces alignment between the resulting current and selected reference expert outputs. Together, OSE provides the directional mechanism for residual exploration in expert-output space, while AMMD localizes the representation components through which the direct OSE signal is applied.

D Additional Analysis

D.1 Redundant Expert Learning Analysis

We analyze redundant expert learning through the representation overlap between the current and reference experts. We examine two aspects: whether stronger historical relevance is associated with greater representation overlap under historical reuse, and whether OSE reduces this overlap.



(a) Relevance-dependent representation redundancy.

(b) Absolute representation redundancy.

Figure 5: **Analysis of redundant expert learning and the effect of OSE.** (a) Representation overlap between the current and selected reference experts across five reference-score quintiles, normalized to the Q1 mean for each method. Q1 and Q5 correspond to samples with the lowest and highest selected reference scores, respectively, and percentages indicate the relative change from Q1 to Q5. Spearman's r_s and the corresponding p-values are computed over individual held-out samples. (b) Raw representation overlap g across the same reference-score quintiles on a logarithmic scale. CORE exhibits consistently lower current/reference expert overlap across all quintiles.

D.1.1 Evidence of Redundant Expert Learning

To examine the relationship between historical relevance and current/reference expert representation overlap, we analyze the **+Reuse** variant, which retains CORE's reference scoring and historical expert reuse without OSE. This setting allows us to characterize current/reference expert overlap under historical reuse before OSE is applied. For each held-out sample x_i^t, we obtain its selected reference score $s_i^{\star,t}$ and measure the layer-averaged squared normalized alignment between the current and selected reference expert outputs,

$$g_i^t = \frac{1}{L}\sum_{\ell=1}^{L} \frac{\left((u_i^{t,\ell})^\top u_i^{\star,t,\ell}\right)^2}{\|u_i^{t,\ell}\|_2^2 \|u_i^{\star,t,\ell}\|_2^2}. \tag{74}$$

A larger g_i^t indicates stronger directional overlap between the current and selected reference expert representations. We sort held-out samples by $s_i^{\star,t}$, divide them into five equally sized quintiles, and report the mean representation overlap within each quintile. We further compute Spearman's rank correlation r_s between $s_i^{\star,t}$ and g_i^t over individual held-out samples.

As shown in Figure 5a, the mean representation overlap increases by approximately 18% from Q1 to Q5 under historical reuse. At the sample level, the selected reference score is positively associated with representation overlap ($r_s = 0.17$, $p = 2 \times 10^{-5}$). Together, these results show that samples with stronger historical relevance tend to exhibit greater overlap between the current and selected reference experts. This relevance-dependent increase in representation overlap provides empirical evidence for redundant expert learning under historical reuse and motivates learning the current expert as a residual relative to relevant historical experts.

D.1.2 Effect of OSE on Representation Redundancy

We next examine whether OSE effectively reduces the representation overlap identified above. Using the same held-out samples and overlap metric g_i^t, we compare **+Reuse** with **CORE** across the five reference-score quintiles. As shown in Figure 5b, CORE consistently produces substantially lower current/reference expert overlap across all quintiles. Specifically, the overlap is reduced by 9.1×, 9.1×, 9.6×, 8.5×, and 10.4× from Q1 to Q5, respectively, corresponding to an average reduction of approximately 9.3×. At the sample level, 97.2% of held-out samples exhibit lower representation overlap under CORE, and the paired Wilcoxon test shows that the reduction is statistically significant ($p \approx 10^{-104}$).

The relevance-dependent pattern is also markedly weaker under CORE. As shown in Figure 5a, the Q1-to-Q5 increase decreases from approximately 18% under **+Reuse** to 3% under **CORE**, while the sample-level Spearman correlation decreases from $r_s = 0.17$ to $r_s = 0.01$. Together, these results show that OSE not only reduces current/reference expert overlap by nearly an order of magnitude, but also largely attenuates the relevance-dependent increase in overlap observed under historical reuse. This provides direct empirical support for the role of OSE in guiding the current expert toward complementary residual representations relative to relevant historical experts.

D.2 Reference Scoring and Reference Expert Selection

CORE uses reference scores to govern both historical expert reuse and reference expert selection. We therefore examine two aspects: whether sample-level reference selection is more effective than alternative reference-selection strategies under fixed historical reuse, and whether dual-modal scoring provides a more effective relevance signal than either image-only or instruction-only scoring.

D.2.1 Reference Expert Selection under Fixed Historical Reuse

To isolate the effect of reference expert selection, we keep historical reuse unchanged across variants by using the same dual-modal reference scores and softmax allocation weights as CORE, and vary only the reference-selection rule, with AMMD and OSE following the selected expert. We additionally include Reuse Only, which disables reference-guided AMMD and OSE while retaining the same historical reuse, as a no-reference control. We compare four reference-selection strategies: **Random**, which uniformly assigns one historical expert to each sample, keeps the assignment fixed throughout training, and reports the mean over three independent assignment seeds; **Latest**, which always selects the most recently learned expert; **Task-Level**, which selects a single historical expert with the highest average reference score over the current-task training samples; and **Sample-Level (CORE)**, which selects the highest-scoring historical expert independently for each sample. For each strategy, the OSE weight is determined by the selected expert's reference score, while all other training and inference settings remain unchanged. We report the final accuracies on all eight CoIN tasks together with MFN, MAA, and BWT.

Table 4: Reference expert selection under fixed historical reuse on CoIN.

Ref Strategy	Final Accuracy on Each Task (%)								Aggregate Metrics (%)		
	SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA^{v2}	VQA^{OCR}	MFN↑	MAA↑	BWT↑
Reuse Only	76.59	51.06	96.33	47.54	44.60	6.74	57.44	60.57	55.11	66.85	−7.38
Random	76.42	50.73	96.29	47.18	44.11	6.49	57.21	59.84	54.78	66.48	−7.81
Latest	76.38	51.37	96.21	47.93	45.41	7.02	57.86	60.39	55.32	66.94	−7.14
Task-Level	76.88	55.48	96.47	50.59	52.38	12.74	60.12	62.41	58.38	68.29	−4.91
Sample-Level	**77.75**	**57.24**	**96.73**	**52.41**	**55.98**	**15.21**	**62.15**	**63.71**	**60.15**	**69.29**	**−3.83**

Table 4 shows that arbitrary or recency-based reference selection provides little benefit beyond historical reuse alone, with Random and Latest remaining close to Reuse Only. In contrast, Task-Level reference selection improves MFN, MAA, and BWT by 3.27, 1.44, and 2.47 points over Reuse Only, respectively, demonstrating the benefit of selecting a relevant historical reference. Sample-Level selection further improves over Task-Level by 1.77, 1.00, and 1.08 points in MFN, MAA, and BWT,

while achieving higher final accuracy on all eight tasks. These results suggest that sample-level reference selection better accounts for sample-specific variation in historical relevance within a task, supporting this sample-adaptive strategy in CORE.

D.2.2 Reference Scoring Mechanism

To examine the reference scoring mechanism, we compare three reference-score formulations: **Image-Only**, which uses only image similarity; **Instruction-Only**, which uses only instruction similarity; and **Dual-Modal (CORE)**, which averages the two modality-specific similarities. For each variant, the score is used consistently to compute allocation weights, select the reference expert, and determine the OSE weight, while all other training and inference settings remain unchanged. We report the final accuracies on all eight CoIN tasks together with MFN, MAA, and BWT.

Table 5: Comparison of reference-score formulations on CoIN.

Ref Score	Final Accuracy on Each Task (%)								Aggregate Metrics (%)		
	SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA^{v2}	VQA^{OCR}	MFN↑	MAA↑	BWT↑
Instruction-Only	77.05	56.20	96.30	50.55	52.10	12.70	60.25	62.35	58.44	68.21	−5.01
Image-Only	77.10	55.64	96.60	51.20	53.40	13.70	61.10	61.82	58.82	68.48	−4.62
Dual-Modal	**77.75**	**57.24**	**96.73**	**52.41**	**55.98**	**15.21**	**62.15**	**63.71**	**60.15**	**69.29**	**−3.83**

Table 5 shows that Image-Only and Instruction-Only exhibit different task-level strengths, with Image-Only performing better on six tasks while Instruction-Only performs better on TextVQA and OCR-VQA. Dual-Modal scoring outperforms both unimodal variants, achieving the highest final accuracy on all eight tasks. Compared with the stronger unimodal variant in aggregate performance, Image-Only, Dual-Modal improves MFN, MAA, and BWT by 1.33, 0.81, and 0.79 points, respectively. These results indicate that image and instruction similarities provide complementary signals for estimating historical relevance, supporting their joint use for reference scoring in CORE.

D.3 AMMD and Image-Centric Head Analysis

AMMD uses task-specific image-centric head sets to localize residual exploration to visual-oriented representations. We therefore evaluate sensitivity to the head-selection threshold and analyze how image-centric head patterns vary across transformer layers and tasks.

D.3.1 Head-Selection Threshold Sensitivity

To evaluate the sensitivity of AMMD to the image-centric head-selection threshold, we vary ρ over $\{0.6, 0.7, 0.8, 0.9\}$, where 0.8 is the default setting. All other components and training settings remain unchanged. For each threshold, we report MFN, MAA, and BWT.

Table 6: Sensitivity to the image-centric head-selection threshold ρ on CoIN (%).

ρ	MFN ↑	MAA ↑	BWT ↑
0.6	58.91	68.54	−4.67
0.7	59.72	68.96	−4.15
0.8	**60.15**	**69.29**	**−3.83**
0.9	59.34	68.71	−4.31

Table 6 shows that performance improves as ρ increases from 0.6 to 0.8 across all three metrics, while a further increase to 0.9 leads to a decline. The default setting $\rho = 0.8$ achieves the best MFN, MAA, and BWT of 60.15, 69.29, and -3.83, respectively. Performance remains reasonably stable across the tested thresholds, indicating limited sensitivity to the precise threshold choice. A lower threshold retains a broader set of heads, potentially including heads with weaker visual orientation, whereas an overly high threshold may exclude useful heads that still contribute visual information. The results therefore suggest that $\rho = 0.8$ provides a suitable balance between focused head selection and preservation of useful visual information.

D.3.2 Layer-wise Image-Centric Head Distributions

Figure 6 provides a finer-grained view of the image-centric head patterns summarized in Figure 4(c), visualizing visual non-sink scores for individual attention heads across transformer layers and CoIN tasks. Higher scores indicate that a larger proportion of visual attention is assigned to visual non-sink tokens, which AMMD uses to identify image-centric heads. The first two layers exhibit scores close to one across tasks because visual sink tokens have not yet emerged at these early layers. Beyond this shared early-layer pattern, the layer-by-head score patterns differ across tasks, indicating task-dependent visual attention distributions. These task-dependent distributions further support maintaining a task-specific image-centric head set for each historical expert rather than using a fixed head set shared across tasks.

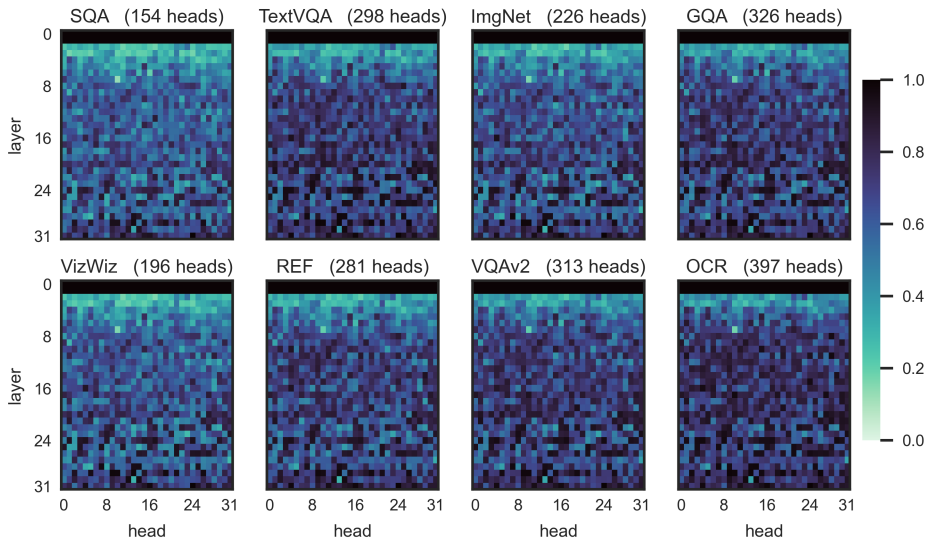


Figure 6: Visual non-sink scores across attention heads for the eight CoIN tasks. Rows correspond to transformer layers and columns to attention heads. Higher scores indicate a larger proportion of visual attention assigned to visual non-sink tokens. Scores in the first two layers are close to one because visual sink tokens have not yet emerged.

D.4 Robustness to Task Order

Following LLaVA-DyMoE (Zhao et al., 2026), we evaluate CORE under three task orders on CoIN: **Origin**, **Reverse**, and **Alphabet**. Origin follows the default CoIN sequence, Reverse uses the reversed sequence, and Alphabet orders the eight tasks alphabetically. All other training and evaluation settings are kept unchanged, and we report MFN, MAA, and BWT for each task order.

Table 7: Robustness of CORE to different task orders on CoIN (%).

Task Order	MFN↑	MAA↑	BWT↑
Origin	60.15	69.29	−3.83
Reverse	59.67	68.84	−4.12
Alphabet	59.07	68.14	−4.47

Table 7 shows that CORE maintains consistently strong performance across all three task orders. While Origin achieves the best overall results, Reverse and Alphabet remain comparable, with only modest variations in MFN, MAA, and BWT. These results demonstrate the robustness of CORE to changes in task order.

D.5 Robustness to Distinct Instruction Templates

Because CORE incorporates instruction similarity into its reference scoring, we further evaluate its robustness to variations in instruction templates. Following LLaVA-DyMoE (Zhao et al., 2026), we adopt the same three instruction-template settings on CoIN: **Original**, **Diverse**, and **10Type**. Original uses the default CoIN instruction templates, whereas Diverse and 10Type use the corresponding alternative templates used in LLaVA-DyMoE. The exact templates used for all eight tasks are provided in Table 16. All other training and evaluation settings, including the default task order, remain unchanged, and we report MFN, MAA, and BWT under each setting.

Table 8: Robustness of CORE to different instruction-template settings on CoIN (%).

Instruction Template	**MFN↑**	**MAA↑**	**BWT↑**
Original	60.15	69.29	−3.83
Diverse	59.88	69.11	−4.06
10Type	59.57	68.83	−4.26

Table 8 shows that CORE maintains consistently strong performance across all three instruction-template settings. The Diverse and 10Type settings yield performance comparable to Original, with only modest variations in MFN, MAA, and BWT. These results demonstrate that CORE remains robust to variations in instruction templates.

D.6 Different Model Scales

To assess whether CORE remains effective across model scales, we extend the evaluation from 7B to 13B. We compare CORE with IncLoRA and LLaVA-DyMoE at both scales, using the 7B results from the main experiments and evaluating the 13B variants under matched training and evaluation settings. We use the default CoIN task order and Original instruction templates, and report task-wise final accuracy together with MFN, MAA, and BWT for each model scale.

Table 9: Performance comparison across different model scales on CoIN (%).

Model Scale	Method	Final Accuracy on Each Task (%)								Aggregate Metrics (%)		
		SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA$^{\text{v2}}$	VQA^{OCR}	MFN↑	MAA↑	BWT↑
7B	IncLoRA	73.33	44.32	54.59	44.07	25.93	4.49	54.91	58.55	45.02	60.73	−16.12
	LLaVA-DyMoE	76.25	53.86	95.80	48.40	52.35	9.25	58.30	62.00	57.03	67.56	−6.04
	CORE	**77.75**	**57.24**	**96.73**	**52.41**	**55.98**	**15.21**	**62.15**	**63.71**	**60.15**	**69.29**	**−3.83**
13B	IncLoRA	74.27	45.88	66.34	46.57	26.76	7.23	57.87	61.96	48.36	63.43	−15.19
	LLaVA-DyMoE	78.75	56.24	96.05	55.85	53.20	13.85	64.05	65.15	60.39	70.06	−5.58
	CORE	**79.91**	**58.84**	**96.69**	**59.13**	**55.96**	**18.24**	**67.12**	**66.49**	**62.80**	**71.25**	**−3.60**

Table 9 shows that all three methods achieve stronger aggregate performance with the 13B backbone than with the 7B backbone. CORE consistently outperforms IncLoRA and LLaVA-DyMoE at both model scales and remains superior to LLaVA-DyMoE across all eight tasks with the 13B backbone. These results demonstrate that the advantage of CORE is maintained when scaling the backbone from 7B to 13B.

D.7 Compatibility with Data-Based Continual Learning Strategies

To assess whether CORE remains effective when combined with data-based continual learning strategies, we consider two representative strategies: Answer Style Diversification (ASD) and replay. For ASD, we follow the data setting of SEFE and compare CORE with SEFE and LLaVA-DyMoE under the same ASD setting (Chen et al., 2025; Zhao et al., 2026). For replay, we use a replay budget of 200 samples in the main comparison, following ProgLoRA and LLaVA-DyMoE, and compare ProgLoRA, LLaVA-DyMoE, and CORE under the matched replay setting (Yu et al., 2025; Zhao et al., 2026). We additionally include MoELoRA and O-LoRA as references without

additional data-based strategies. All reported results correspond to the LLaVA-v1.5-7B backbone, the default CoIN task order, and the Original instruction-template setting. We report task-wise final accuracy together with MFN, MAA, and BWT in the main compatibility comparison, and further evaluate replay budgets of 200, 500, and 1000 samples to examine the effect of replay buffer size.

Table 10: Compatibility with data-based continual learning strategies on CoIN (%). The replay comparison uses a budget of 200 samples.

Data Strategy	Method	Final Accuracy on Each Task (%)								Aggregate Metrics (%)		
		SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA$^{\text{v2}}$	VQA^{OCR}	MFN↑	MAA↑	BWT↑
None	MoELoRA	72.01	46.89	44.75	42.79	28.22	3.31	55.74	57.72	43.93	60.26	−17.11
	O-LoRA	75.61	49.98	78.24	44.18	30.70	4.66	55.51	57.37	49.53	63.58	−10.98
	LLaVA-DyMoE	76.25	53.86	95.80	48.40	52.35	9.25	58.30	62.00	57.03	67.56	−6.04
	CORE	**77.75**	**57.24**	**96.73**	**52.41**	**55.98**	**15.21**	**62.15**	**63.71**	**60.15**	**69.29**	**−3.83**
+ ASD	SEFE	75.35	58.66	83.10	54.25	48.85	16.75	65.35	66.25	58.57	67.66	−5.34
	LLaVA-DyMoE	74.60	55.24	93.80	53.45	55.00	25.50	63.95	62.85	60.55	68.53	−2.96
	CORE	**77.58**	**59.13**	**95.82**	**55.17**	**56.73**	**25.67**	**66.18**	**66.47**	**62.84**	**70.00**	**−1.51**
+ Replay	ProgLoRA	74.84	51.83	83.90	49.93	53.87	31.19	62.71	64.44	59.09	62.40	−6.59
	LLaVA-DyMoE	75.55	56.88	96.50	54.75	55.90	29.15	63.65	64.25	62.08	68.60	−1.85
	CORE	**78.17**	**58.53**	**96.74**	**55.27**	**56.88**	**33.14**	**63.82**	**64.57**	**63.39**	**70.52**	**−1.49**

Compatibility with Data-Based Strategies. Table 10 shows that CORE remains effective when combined with both ASD and replay. Under ASD, CORE achieves an MFN of 62.84%, an MAA of 70.00%, and a BWT of -1.51%, outperforming LLaVA-DyMoE by 2.29, 1.47, and 1.45 percentage points, respectively. With a replay budget of 200 samples, CORE further reaches 63.39% MFN, 70.52% MAA, and -1.49% BWT, exceeding LLaVA-DyMoE by 1.31, 1.92, and 0.36 percentage points, respectively. CORE also achieves the best final accuracy across all eight tasks under both data-based strategies. Compared with CORE without additional data, both ASD and replay further improve the aggregate metrics, indicating that these data-based strategies provide complementary benefits while preserving the advantage of CORE over competing methods.

Table 11: Performance under different replay budgets on CoIN (%).

Replay Budget	Method	MFN↑	MAA↑	BWT↑
200	ProgLoRA	59.09	62.40	−6.59
	LLaVA-DyMoE	62.08	68.60	−1.85
	CORE	**63.39**	**70.52**	**−1.49**
500	ProgLoRA	59.14	62.74	−6.47
	LLaVA-DyMoE	62.55	68.85	−1.28
	CORE	**63.79**	**70.60**	**−0.97**
1000	ProgLoRA	59.66	63.23	−6.21
	LLaVA-DyMoE	63.19	69.30	−0.85
	CORE	**64.10**	**70.70**	**−0.53**

Effect of Replay Buffer Size. Table 11 further evaluates CORE under different replay budgets. As the replay budget increases from 200 to 500 and then to 1000 samples, CORE improves from 63.39% to 63.79% and 64.10% in MFN, from 70.52% to 70.60% and 70.70% in MAA, and from -1.49% to -0.97% and -0.53% in BWT, respectively. CORE consistently achieves the best MFN, MAA, and BWT at every replay budget, while maintaining positive margins over LLaVA-DyMoE and ProgLoRA. These results show that the compatibility of CORE with replay remains robust across different replay budgets.

D.8 Analysis of Historical Expert Reuse Strategies

To further examine how historical experts should be reused during continual training, we compare several alternative aggregation strategies while keeping sample-level reference expert selection, AMMD, OSE, and CORE's standard sample-level expert allocation at final inference fixed. Specifically, all variants use the same dual-modal reference scores for sample-level reference expert selection and differ only in how historical expert outputs are aggregated during current-task training. We consider four strategies: **Uniform**, which assigns equal weights to all historical experts; **Hard Top-1**, which reuses only the highest-scoring historical expert for each sample; **Task-Level Soft**, which averages the reference scores over all training samples of the current task, applies a softmax over historical experts to obtain a single task-level allocation vector, and shares this allocation across samples in that task; and **Sample-Level Soft (CORE)**, which computes relevance-weighted allocation independently for each sample. All other training and evaluation settings remain unchanged. We report the current-task accuracy immediately after each training stage to examine how different reuse strategies affect current-task adaptation across the task sequence, together with task-wise final accuracy, MFN, MAA, and BWT to evaluate overall continual-learning performance.

Table 12: Immediate current-task accuracy under historical expert reuse strategies on CoIN (%).

Reuse Strategy	SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA^{v2}	VQA^{OCR}
Uniform	80.70	60.53	96.72	55.83	55.92	29.73	64.48	62.99
Hard Top-1	80.70	60.53	96.87	56.17	55.31	31.79	64.11	63.48
Task-Level Soft	80.70	60.53	96.81	56.49	55.76	31.28	64.67	63.27
Sample-Level Soft (CORE)	80.70	60.53	**96.95**	**56.77**	**56.13**	**32.11**	**64.95**	**63.71**

Table 13: Final continual-learning performance under historical expert reuse strategies on CoIN (%).

Reuse Strategy	Final Accuracy on Each Task (%)								Aggregate Metrics (%)		
	SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA^{v2}	VQA^{OCR}	MFN↑	MAA↑	BWT↑
Uniform	77.14	55.92	96.40	50.62	53.26	11.64	60.43	62.99	58.55	68.73	−4.81
Hard Top-1	77.31	56.28	96.52	51.07	54.32	12.81	60.89	63.48	59.09	68.94	−4.54
Task-Level Soft	77.48	56.61	96.59	51.48	54.99	13.66	61.44	63.27	59.44	69.07	−4.25
CORE	**77.75**	**57.24**	**96.73**	**52.41**	**55.98**	**15.21**	**62.15**	**63.71**	**60.15**	**69.29**	**−3.83**

Table 12 shows that the relative strengths of alternative reuse strategies vary across tasks. On the first two tasks, all strategies yield identical immediate accuracy because no historical expert is available for SQA and only one historical expert is available for TextVQA. From the third task onward, however, the strongest alternative differs across tasks: Hard Top-1 performs best among the alternatives on ImgNet, REF, and OCR-VQA, Task-Level Soft performs best on GQA and VQAv2, while Uniform performs best on VizWiz. CORE achieves the highest immediate current-task accuracy on every task from T3 to T8, with gains of 0.08–0.32 percentage points over the strongest alternative. Under the controlled setting, the comparison with Hard Top-1 indicates that restricting reuse to only the highest-scoring historical expert is less effective than using soft relevance-weighted aggregation over multiple historical experts, while the comparison with Task-Level Soft indicates that sharing a single soft allocation across an entire task is less effective than adapting the allocation to individual samples. Table 13 further shows that these differences are reflected in stronger overall continual-learning performance. Compared with Task-Level Soft, the strongest alternative in aggregate performance, CORE improves MFN, MAA, and BWT from 59.44%, 69.07%, and -4.25% to 60.15%, 69.29%, and -3.83%, corresponding to gains of 0.71, 0.22, and 0.42 percentage points, respectively. CORE also achieves the highest final accuracy on all eight tasks, supporting sample-level soft allocation as a more effective historical reuse strategy than the evaluated uniform, hard single-expert, and task-level shared aggregation strategies.

D.9 Final-Stage Expert Allocation Analysis

Complementing the training-time reuse analysis above, we further examine the expert-allocation behavior of CORE at final inference, after the expert pool has fully expanded. Specifically, we analyze the final model after learning all T tasks, using the complete expert pool $\mathcal{E}_T = \{E_1, \ldots, E_T\}$. For each test sample x from task k, CORE follows its standard inference procedure and computes a soft allocation weight $a_e(x)$ for each expert E_e, where $e = 1, \ldots, T$.

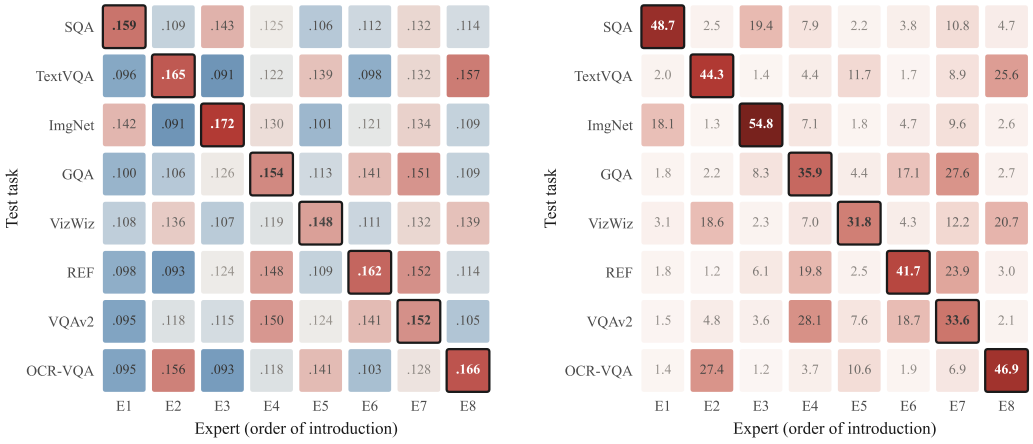
We characterize the final-stage allocation from two complementary perspectives. First, we compute the **mean allocation weight**

$$\bar{a}_{k,e} = \frac{1}{|\mathcal{D}_k^{\text{test}}|} \sum_{x \in \mathcal{D}_k^{\text{test}}} a_e(x), \tag{75}$$

which measures the average soft allocation assigned to expert E_e over the test samples of task k. Second, we compute the **Top-1 expert frequency**

$$p_{k,e} = \frac{1}{|\mathcal{D}_k^{\text{test}}|} \sum_{x \in \mathcal{D}_k^{\text{test}}} \mathbb{I}\left[e = \underset{j \in \{1,\ldots,T\}}{\arg\max}\ a_j(x) \right], \tag{76}$$

which measures the proportion of task-k test samples for which E_e receives the largest allocation weight. We visualize $\bar{a}_{k,e}$ and $p_{k,e}$ as task-by-expert heatmaps, with rows corresponding to test tasks and columns corresponding to experts in their order of introduction. Together, the two views characterize the average soft contribution of each expert and the distribution of dominant experts across individual samples after continual expansion.



(a) Mean allocation weight.

(b) Top-1 expert frequency.

Figure 7: **Final-stage expert allocation of CORE after training on all CoIN tasks.** (a) Mean soft allocation weight assigned to each expert across the test samples of each task. (b) Top-1 expert frequency, i.e., the percentage of test samples for which each expert receives the largest allocation weight. Rows denote test tasks and columns denote experts in their order of introduction. Black boxes indicate the expert introduced for the corresponding task.

Figure 7 shows that the final-stage expert allocation remains structured after the expert pool has fully expanded. In both panels, the expert introduced for each task attains the row-wise maximum for all eight tasks. The diagonal mean allocation weights range from 0.148 to 0.172, while the diagonal Top-1 frequencies range from 31.8% to 54.8%. This consistent row-wise pattern indicates that the final allocation retains a clear task-associated structure even when all learned experts are simultaneously available. At the same time, the allocation remains distributed across multiple experts rather than being concentrated exclusively on the expert introduced for each task. Off-diagonal allocations are also observed for several task pairs. For example, TextVQA and OCR-VQA assign relatively high weights to each other's experts, while a similar cross-task pattern appears between GQA and

VQAv2. The Top-1 frequencies further show that off-diagonal experts are dominant for a subset of samples, indicating that final-stage allocation is not restricted to the expert introduced for each task. Moreover, recently introduced experts do not systematically dominate earlier tasks at final inference.

E QUALITATIVE RESULT EXAMPLES



Figure 8: **Qualitative comparison between CORE and IncLoRA on CoIN.** Examples are shown from ScienceQA, ImageNet, TextVQA, and GQA. Each example shows the image, question, predictions of IncLoRA and CORE, and the ground-truth answer. Check marks and crosses indicate correct and incorrect predictions, respectively.

Figure 8 presents qualitative comparisons between CORE and IncLoRA across four CoIN tasks spanning science question answering, image recognition, scene-text understanding, and composi-

tional visual question answering. Across these examples, CORE corrects several errors made by IncLoRA. For example, CORE correctly identifies the highlighted continent in ScienceQA, recognizes the camera brand in TextVQA, and more accurately identifies visual attributes and objects in GQA. Similar improvements are observed for fine-grained visual recognition on ImageNet. While some challenging cases remain difficult for both methods, CORE overall produces more accurate predictions across the illustrated examples. These qualitative results complement the quantitative evaluation by showing that the improvements of CORE extend across diverse multimodal task types.

F Efficiency Analysis

We evaluate the computational overhead of CORE through a controlled comparison with IncLoRA, which follows the same expanding LoRA paradigm but does not include the reference-guided reuse, AMMD, and OSE components introduced by CORE. This controlled comparison more directly reflects the additional computation introduced by CORE-specific components. We report both continual-training time and final-stage inference latency under matched settings for the two methods.

For training-time measurement, both methods use the same continual-learning configuration on two NVIDIA H200 GPUs, and wall-clock adaptation time is recorded separately for each task with CUDA synchronization at the timing boundaries. The timing for CORE includes computations required during continual training, including reference-related computation, historical expert reuse, AMMD and OSE, and the construction of dual-modal anchors and image-centric head sets, while unrelated overhead such as model loading, checkpoint saving, and final evaluation is excluded. For inference latency, we evaluate the models after all eight tasks on a single NVIDIA H200 GPU under identical decoding settings, using batch size 1 and greedy decoding with a fixed generation length of 32 new tokens. CUDA synchronization is applied at the timing boundaries, and the latency of CORE includes the inference pipeline, including online image and instruction encoding, reference-score computation, allocation-weight computation, expert aggregation, and MLLM generation.

Table 14: Training time of CORE and IncLoRA across the eight-task CoIN continual-learning sequence. Per-task, average, and total wall-clock training times are reported in minutes.

Method	Training Time on Each Task (min)								Average	Total
	SQA	VQA^{Text}	ImgNet	GQA	VizWiz	REF	VQA^{v2}	VQA^{OCR}		
IncLoRA	9.2	31.9	124.1	856.2	21.4	343.9	454.7	865.0	338.3	2706.4
CORE	9.3	36.1	133.0	939.8	24.4	378.4	505.5	958.8	373.2	2985.3

Table 14 shows that CORE introduces an increase in training time compared with IncLoRA. Over the full eight-task continual-learning sequence, the total training time increases from 2706.4 to 2985.3 minutes, corresponding to an overhead of 10.31%, while the average training time per task increases from 338.3 to 373.2 minutes. Relative to IncLoRA, CORE improves MFN, MAA, and BWT by 15.13, 8.56, and 12.29 percentage points, respectively, in the main evaluation. These results indicate that the performance gains of CORE are achieved with an approximately 10% training-time overhead, suggesting a favorable efficiency–performance trade-off under the evaluated setting.

Table 15: Final-stage inference latency of CORE and IncLoRA after learning all eight CoIN tasks. Latency is reported in seconds per sample, with overhead measured relative to IncLoRA.

Method	Latency (s/sample)	Overhead (%)
IncLoRA	2.573	–
CORE	2.868	+11.47

Table 15 shows that CORE introduces an increase in inference latency compared with IncLoRA. The mean latency increases from 2.573 to 2.868 seconds per sample, corresponding to an overhead of 11.47%. Because the measurement includes the CORE inference pipeline described above, this overhead reflects the end-to-end cost of the CORE-specific inference operations rather than an isolated component. Together with the continual-learning gains reported in the main evaluation, these results suggest a favorable efficiency–performance trade-off at inference under the evaluated setting.

Table 16: Instruction templates used for the instruction-template robustness evaluation on CoIN.

Task	Original	Diverse	10Type
SQA	Answer with the option's letter from the given choices directly	Answer with the option's letter from the given choices directly	Answer with the option's letter from the given choices directly Select the correct answer from the given choices and respond with the letter of the chosen option Determine the correct option from the provided choices and reply with its corresponding letter Pick the correct answer from the listed options and provide the letter of the selected option Identify the correct choice from the options below and respond with the letter of the correct option From the given choices, choose the correct answer and respond with the letter of that choice Choose the right answer from the options and respond with its letter Select the correct answer from the provided options and reply with the letter associated with it From the given choices, select the correct answer and reply with the letter of the chosen option Identify the correct option from the choices provided and respond with the letter of the correct option From the given choices, pick the correct answer and respond by indicating the letter of the correct option
VQA^{Text}	Answer the question using a single word or phrase	Capture the essence of your response in a single word or a concise phrase	Answer the question with just one word or a brief phrase Use one word or a concise phrase to respond to the question Answer using only one word or a short, descriptive phrase Provide your answer in the form of a single word or a brief phrase Use a single word or a short phrase to respond to the question Summarize your response in one word or a concise phrase Respond to the question using a single word or a brief phrase Provide your answer in one word or a short, descriptive phrase Answer the question with a single word or a brief, descriptive phrase Capture the essence of your response in one word or a short phrase Capture the essence of your response in a single word or a concise phrase
ImgNet	Answer the question using a single word or phrase	Express your answer in a single word or a short, descriptive phrase	Express your answer in a single word or a short, descriptive phrase Provide your answer using a single word or a brief phrase Describe the content of the image using one word or a concise phrase Respond to the question with a single word or a short, descriptive phrase Classify the image content using only one word or a brief phrase Give your answer in the form of a single word or a concise phrase Use a single word or a short phrase to categorize the image content Express your answer with one word or a short, descriptive phrase Identify the type of content in the image using one word or a concise phrase Summarize your response in a single word or a brief phrase Use one word or a short phrase to classify the content of the image
GQA	Answer the question using a single word or phrase	Respond to the question briefly, using only one word or a phrase	Respond to the question with a single word or a short phrase Respond to the question using only one word or a concise phrase Answer the question with a single word or a brief phrase Respond with one word or a short phrase Provide your answer in the form of a single word or a concise phrase Respond to the question with just one word or a brief phrase Answer the question using a single word or a concise phrase Provide your response using only one word or a short phrase Respond to the question with a single word or a brief phrase Respond to the question using just one word or a concise phrase Answer the question with one word or a short phrase
VizWiz	Answer the question using a single word or phrase	Provide a succinct response with a single word or phrase	Answer the question using only one word or a concise phrase Respond to the question using only one word or a concise phrase Respond to the question with a single word or a brief phrase Provide your answer using just one word or a short phrase Respond with one word or a concise phrase Answer the question with just one word or a brief phrase Use a single word or a short phrase to answer the question Provide your answer in the form of one word or a brief phrase Reply to the question using one word or a concise phrase Answer with a single word or a short phrase Use one word or a brief phrase to answer the question
REF	Please provide the bounding box coordinate of the region this sentence describes	Please provide the bounding box coordinate of the region this sentence describes	Identify and provide the bounding box coordinates that match the description given in this sentence Extract and provide the bounding box coordinates based on the region described in the sentence Please provide the bounding box coordinate of the region this sentence describes Find and provide the bounding box coordinates for the region mentioned in the sentence Provide the coordinates of the bounding box that correspond to the region described in the sentence Give the bounding box coordinates as described in the sentence Determine and provide the bounding box coordinates based on the description in the sentence Identify and provide the coordinates of the bounding box described in the sentence Provide the coordinates for the bounding box based on the region described in the sentence Extract and provide the coordinates for the bounding box described in the sentence Identify and give the coordinates of the bounding box as described by the sentence
VQA$^{\text{v2}}$	Answer the question using a single word or phrase	Answer the question using a single word or phrase	Answer the question using a single word or phrase Answer the question with a single word or a brief phrase Use one word or a short phrase to respond to the question Answer the question using just one word or a concise phrase Provide your answer to the question using only one word or a brief phrase Respond to the question with a single word or a short phrase Use a single word or phrase to answer the question Provide an answer using only one word or a brief phrase Answer the question succinctly with one word or a brief phrase Answer the question with just one word or a short phrase Respond to the question using a single word or a concise phrase
VQA^{OCR}	Answer the question using a single word or phrase	Condense your answer for each question into a single word or concise phrase	Answer with the option's letter from the given choices directly Select the correct answer from the given choices and respond with the letter of the chosen option Determine the correct option from the provided choices and reply with its corresponding letter Pick the correct answer from the listed options and provide the letter of the selected option Identify the correct choice from the options below and respond with the letter of the correct option From the given choices, choose the correct answer and respond with the letter of that choice Choose the right answer from the options and respond with its letter Select the correct answer from the provided options and reply with the letter associated with it From the given choices, select the correct answer and reply with the letter of the chosen option Identify the correct option from the choices provided and respond with the letter of the correct option From the given choices, pick the correct answer and respond by indicating the letter of the correct option

G BROADER IMPACT AND LIMITATIONS

This work contributes to the development of more capable and adaptable multimodal continual learning systems by enabling MLLMs to acquire new instruction-following capabilities while effectively retaining previously learned knowledge. CORE further promotes efficient knowledge reuse by leveraging frozen historical LoRA experts and learning complementary residual representations, providing a practical approach to continual adaptation without requiring access to historical task data. This capability can support the long-term adaptation of multimodal models as tasks and user requirements evolve over time. Regarding ethical considerations, CORE builds upon a pretrained LLaVA backbone and standard datasets from the CoIN benchmark, and therefore naturally inherits the characteristics of these underlying models and data. While our current work focuses on balancing knowledge retention and plasticity, future research could further explore safety alignment and fairness as integral components of continual learning for real-world applications.